



A University of Sussex PhD thesis

Available online via Sussex Research Online:

<http://sro.sussex.ac.uk/>

This thesis is protected by copyright which belongs to the author.

This thesis cannot be reproduced or quoted extensively from without first obtaining permission in writing from the Author

The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the Author

When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given

Please visit Sussex Research Online for more information and further details

Synaesthesia in Children

Louisa J Rinaldi

Thesis submitted for the degree of Doctor of Philosophy

School of Psychology

University of Sussex

August 2019

Final revisions March 2020

Declaration

The thesis conforms to an ‘article format’ in which Chapters 2-5 consist of discrete articles written in a style that is appropriate for publication in peer-reviewed journals in the field. Chapter 1 and Chapter 6 present synthesis overviews, and discussions of the field and the research undertaken.

Chapter 2 has been submitted for publication:

Rinaldi, L.J., Smees, R, Carmichael, D. C., & Simner, J (2019) Big Five Personality Instruments for Parents and Children 6+ years: The Pictorial BFI-10-C; the Definitional BFI-44-c, and the BFI-44-parent. Manuscript submitted for publication.

The author contributions are as follows: Louisa Rinaldi was responsible for creating materials and collecting data (with assistance from Rebecca Smees and Duncan Carmichael); analysing data; and writing the manuscript. Julia Simner (supervisor) gave feedback on drafts, and was the grant-holder for the project involved in all stages apart from data collection and analysis.

Chapters 3 has been submitted for publication:

Rinaldi, L.J., Smees, R, Carmichael, D. C., & Simner, J (2019) *What is the personality profile of a child synaesthete?* Manuscript submitted for publication.

The author contributions are as follows: Louisa Rinaldi was responsible for creating materials and collecting data (with assistance from Rebecca Smees and Duncan Carmichael); analysing data; and writing the manuscript. Julia Simner (supervisor) gave feedback on drafts, and was the grant-holder for the project involved in all stages apart from data collection and analysis.

Chapter 4 is in press in *Child Development* as:

Rinaldi, L.J., Smees, R, Alvarez, J, Carmichael, D. C., & Simner, J (*in press*). Do the colors of educational number-tools improve children’s mathematics and numerosity? *Child Development*.

The author contributions are as follows: Louisa Rinaldi was responsible for preparing materials; collecting data (with assistance from Rebecca Smees); analysing data (with assistance from programmer James Alvarez who consulted on analyses and was responsible for programming the Monte Carlo simulation); and writing the manuscript. Julia Simner (supervisor) gave feedback on drafts and was the grant-holder for the project, involved in all stages apart from data collection and analysis.

Chapter 5 has been prepared in paper style for publication:

Rinaldi, L.J., Smees, R, Carmichael, D. C., & Simner, J (2019) *Numeracy Skills in Child Synaesthetes: Evidence from grapheme-colour synaesthesia*. Manuscript in preparation.

The author contributions are as follows: Louisa Rinaldi was responsible for preparing materials; collecting data (with assistance from Rebecca Smees and Duncan Carmichael); analysing data and writing the manuscript. Julia Simner (supervisor) gave feedback on drafts and was the grant-holder for the project, involved in all stages apart from data collection and analysis.

I hereby declare that this thesis has not been and will not be submitted in whole or in part to another University for the award of any other degree.

Louisa J Rinaldi
August 2019

Acknowledgements

I owe thanks to many people who made this thesis possible. First and foremost to Professor Julia Simner for undertaking a huge project and welcoming me into it. Thank you for excellent supervision and guidance, for always improving my work and making me a better researcher. Thanks also to Professor Andy Field for his help, and for agreeing to be my second supervisor even though the main project I wanted his help with I did not end up performing for this thesis.

I am especially grateful to Rebecca Smees, Dr Duncan Carmichael, and Dr Anna Hobbs who worked tirelessly with me on the children's MULTISENSE project. To Rebecca and Duncan for getting schools on board, helping to operationalise the project and for working as a great team when testing in schools. To Anna for the huge amount of admin involved in contacting schools and arranging our testing schedule. To the rest of the MULTISENSE lab, thanks for always being supportive I have learnt a lot from you all.

I would also like to thank everyone else who helped us to test, and who helped us conquer the mammoth amount of data entry, especially to Juliet Wilkes, Mira Kaut, Olivia Casey-Haworth, Monica Boutros, Dermot Crowley, and Molly Berenhaus.

Thanks to the 22 schools who were involved, and who very patiently answered our (many) questions, and let us test in their schools, and thanks to the children who completed everything we asked them to do even when we asked them to do strange things like pair letters with faces.

Thanks to all my friends for being with me throughout and providing necessary distractions. Thanks to my family for pushing me out of my comfort zone, I don't think moving across the country was really what you had in mind but it has turned out for the best.

Finally, to Jennifer for reading, listening, and giving advice throughout. Thanks also for just generally being there and keeping life fun.

Summary

Synaesthesia is a developmental condition that triggers phantom sensations (e.g., colours or tastes) when exposed to everyday stimuli such as graphemes, music, and pain. Yet, despite synaesthesia being a developmental condition, there is very little work in children to date. In this thesis, I explore two types of synaesthesia in children aged 6-10 years old; grapheme-colour synaesthesia (letters and numbers elicit colours) and grapheme-personality synaesthesia (letters and numbers elicit personalities). I first use tests designed specifically for children to identify individuals with these types of synaesthesia. Here I tested children with and without synaesthesia who had been identified from a very large screening endeavour, called MULTISENSE (funded by the European Research Council; I played a central role in this project, but my thesis focuses on the children identified by this process, rather than the screening itself). Then once this cohort was identified I looked at group differences between synaesthetes and non-synaesthetes in two domains: personality and cognition (specifically, numerical cognition). Throughout the thesis I use tests targeted specifically for our child population. Where these did not already exist in Chapter 2 (e.g., suitable self-report personality measures for children) we created and validated them independently. In Chapter 3 I use some of these measures to identify whether synaesthetes have a different personality profile to non-synaesthetes. In the second half of the thesis I tested synaesthetes' numerical cognition, and looked, too, at 'synaesthesia-like' phenomena in the general population. Here in Chapter 4 I explored whether a widely implemented maths tool that pairs numbers with colours aids non-synaesthete children in their numerical cognition. I then finally return to synaesthetes in Chapter 5 using the same tests of numerical cognition to determine if grapheme-colour synaesthetes show advantages in this domain. Overall, this thesis shows that child synaesthetes have a distinct personality profile, and show a pattern of differences in numerical cognition found also in 'synaesthesia-like' phenomena such as the educational colour-coding of numbers.

Table of Contents

Declaration	2
Acknowledgements.....	4
Summary	5
Chapter 1	
General Introduction	8
Identifying synaesthetes	10
Letter and Number acquisition.....	13
Synaesthesia in children	14
Identifying synaesthesia in children: The MULTISENSE project	17
Environmental synaesthesia and unusual associations in non-synaesthetes	19
Overview of experimental chapters.....	20
Summary	23
Chapter 2	
Big Five Personality Instruments for Parents and Children 6+ years: The Pictorial BFI-10-C; the Definitional BFI-44-C, and the BFI-44-Parent.	24
Chapter Summary.....	24
Abstract.....	25
Introduction.....	26
Three New and/ or Newly-validated Instruments	29
Study 1: Validating the Big Five Inventory-44-Parent	31
Methods.	31
Results	33
Discussion.....	37
Study 2: Validating children's self-report measures; The Definitional BFI-44-C and the Pictorial BFI-10-C.....	38
Methods	38
Results.	46
Discussion.....	57
General Discussion.....	58
Chapter 2: Supplementary Information	62
Chapter 3	
What is the personality profile of a child synaesthete?.....	64
Chapter Summary.....	64
Abstract.....	65
Introduction.....	66
Methods	71

Results	75
Discussion.....	88
Chapter 3: Supplementary Information	94
Chapter 4	
Do the colors of educational number-tools improve children's mathematics and numerosity?	98
Chapter Summary.....	98
Abstract.....	99
Introduction.....	100
Methods	107
Results	111
Discussion.....	119
Chapter 4: Supplementary Information	125
Chapter 5	
Numeracy Skills in Child Synaesthetes: Evidence from grapheme-colour synaesthesia	129
Chapter Summary.....	129
Abstract.....	130
Introduction.....	131
Methods	136
Results	140
Discussion.....	145
Chapter 5: Supplementary Information	149
Chapter 6	
General Discussion	152
Personality: What we have learnt?.....	152
Numerical cognition: What we have learnt?.....	156
Other future directions.....	159
Designing and running tests for a large cohort of children	161
Conclusions.....	163
References	165
Appendices	185
Appendix A.....	185
Appendix B	190
Appendix C	194
Appendix D.....	197
Appendix E	198
Appendix F	199

Chapter 1

General Introduction

A small percentage of the population have a different perception of the world. These individuals experience colours, smells or even tastes during everyday activities like reading or talking. Synaesthesia is a developmental condition and affects at least 4.4% of the population (Simner, Mulvenna, et al., 2006) which, whilst a small percentage, is still approximately 338 million people worldwide. In this thesis I investigate synaesthesia in children in order to better understand those children who can experience it, and how it impacts them.

One of the types of synaesthesia I focus on in this thesis is *grapheme-colour synaesthesia* -- in which letters and numbers are perceived to have colours (Baron-Cohen, Wyke, & Binnie, 1987; Simner, Mulvenna, et al., 2006). This is one of the more common types of synaesthesia, estimated to occur in approximately 1.2% of the population in both adults and children (Carmichael, Down, Shillcock, Eagleman, & Simner, 2015; Simner, Harrold, Creed, Monro, & Foulkes, 2009; Simner, Mulvenna, et al., 2006). Melanie Ahrlin, is a grapheme-colour synaesthete who gave a personal account of her synaesthesia in a book written in 2009 (Dittmar, 2009). She describes her synaesthesia as follows: “I actually see an ‘A’ set in red before me on the page, I hear it in red and in my mind it is red” (Dittmar, 2009, pg.143). So it is clear that for synaesthetes such as Ahrlin there is an undeniable link between a trigger, referred to as *the inducer* in scientific terminology (Grossenbacher & Lovelace, 2001), and its unusual synaesthetic experience, known as the *concurrent*. Another type of synaesthesia I will focus on in this thesis is *sequence-personality synaesthesia* (also known as *ordinal linguistic personification OLP*¹) in which letters and numbers induce concurrents that are personifications (i.e. personalities and/ or genders; Simner, Glover, & Mowat, 2006; Simner & Holenstein, 2007). For example, Simner and Holenstein (2007) describe an OLP synaesthete for whom “*m* [is an] old lady, like *n*; they

¹ We use this acronym throughout as it is the most widely used name for the condition (1820 hits on google scholar compared to 35 for sequence-personality synaesthesia)

spend all their time together and gossip a lot” (Simner & Holenstein, 2007, pg.696). In this thesis, I will investigate both OLP synaesthesia and grapheme-colour synaesthesia, in primary school aged children.

So why might we want to investigate synaesthesia in children? Synaesthesia research in adults has determined a standardised way of testing for synaesthesia, made important strides towards determining the genetic basis and familial inheritance of synaesthesia, investigated a number of advantages and cognitive differences associated with synaesthesia (see below), and discovered some overlaps and comorbidities with other conditions (see e.g., Baron-Cohen et al., 1987; Carmichael, Smees, Shillcock, & Simner, 2018; Chun & Hupé, 2016; Eagleman, Kagan, Nelson, Sagaram, & Sarma, 2007; Rouw & Scholte, 2016; Rouw, Scholte, & Colizoli, 2011). I will explore more of these findings below but for now, I note that research into the development of the condition during childhood is rare and still largely unexplored. This thesis aims to investigate the development of synaesthesia by looking at child-synaesthetes previously identified from a random sample of over 3000 children. This earlier screening (carried out by the MULTISENSE project, of which this thesis is a product of) served to identify grapheme-colour and sequence-personality synaesthetes, while this thesis itself examines differences in their personality and cognition.

In this thesis I will focus on the cognitive and personality profiles of children with grapheme-colour synaesthesia and OLP synaesthesia, but in order to do so it was first important to consider whether we had adequate childhood tests at our disposal. One component of this thesis will identify an absence of viable personality tests for young children, and will fill that gap with a novel validated test in Chapter 2. In Chapter 3, I will then use this personality test to identify whether child synaesthetes show a particular personality profile. In Chapters 4 and 5 I focus solely on colours (rather than personifications), and again ask whether the unusual cross-modal associations of synaesthetes give them differences beyond synaesthesia itself, this time looking at their cognition. Hence, in Chapter 5 I ask whether grapheme-colour synaesthetes are better in numerical cognition compared to non-synaesthetes. In Chapter 4, I ask a similar question about a group of children who have synaesthesia-like coloured numbers, but are not synaesthetes. Instead, these children have acquired their number-colour associations from using an educational maths tool at school. This tool (see Chapter 4) pairs numbers with colours and some children naturally learn these colour associations. I therefore ask

whether these children show advantages in their numerical cognition, similar to synaesthetes. (Non-synaesthetes are presented in Chapter 4 prior to synaesthetes in Chapter 5, to reflect the way in which these chapters were released into the peer reviewed literature, and therefore cross-reference each other)

Before presenting these findings, I will first summarise the literature on how to identify synaesthesia, and I will also summarise the development of literacy and numeracy; since letters and numbers serve as the synaesthetic inducers for both types of synaesthesia I focus on. I will then summarise the existing research on synaesthesia in children, and where relevant, describe the MULTISENSE project, which this thesis forms part of. I then will summarise some of the literature related to synaesthesia-like associations in non-synaesthetes, and cases of environmentally driven synaesthesia, which ties into Chapter 4 (where I explore environmentally driven coloured number associations in non-synaesthetes).

Identifying synaesthetes

This thesis forms part of a project that identified child synaesthetes using newly developed child-oriented tests. In order to understand the validity of these tests we must understand the foundational literature that led to the first objective identification of synaesthetes in adults. Below I summarise this foundational literature, first looking at grapheme-colour synaesthesia, and then turning to more recent extensions into OLP synaesthesia.

One of the first questions early researchers faced was how to verify that self-declared synaesthetes were truly synaesthetes. A key feature of synaesthesia is the automaticity (Robertson & Sagiv, 2004) and long-term consistency of synaesthetes' associations (Baron-Cohen et al., 1987; Simner & Logie, 2007). For example, a grapheme-colour synaesthete that has a red letter A will likely still have the same colour-association years later (for discussion see; Simner, 2012). This critical feature of synaesthesia was first noted by Baron-Cohen, Wyke and Binnie (1987) who used consistency to objectively verify grapheme-colour synaesthesia for the first time. In their *test of genuineness* they tested a synaesthete and recorded her colour associations for letters and numbers. They asked a control to invent similar associations and then re-tested both synaesthete and a control 3 hours later. They then retested both groups again but tested the synaesthete again

ten weeks later and the control only 2 weeks later, ‘stacking the odds’ in favour of the control. They found the synaesthete to be consistent over time (100% recall) whereas the control was not (17% recall). These findings were replicated in 1993 with a larger group of synaesthetes (Baron-Cohen, Harrison, Goldstein, & Wyke, 1993). In this study Baron-Cohen et al. (1993) again ‘stacked the odds’ in favour of the control participants by giving the synaesthetes a longer interval of a year before re-testing, compared to the control participants who were retested after a week. In this newly-formed diagnostic for genuineness, synaesthetes were required to be more consistent than controls even though synaesthetes had a longer retention interval. This method of assessing consistency as a diagnostic for synaesthesia has become the most widely used *gold standard* test (Rich, Bradshaw, & Mattingley, 2005) in almost all subsequent studies.

This idea of consistency was refined by Ward and Simner, who were the first to automate the consistency test using digital colour palettes (Simner, Mulvenna, et al., 2006; Ward, Huckstep, & Tsakanikos, 2006). Consistency was again later defined by Eagleman, Kagan, Nelson, Sagaram and Sarma (2007) who created an online testing battery to verify synaesthetes in a single testing session (i.e., where test and re-test happened within the same session). In this version of the test, participants see an on-screen colour palette with over 16 million colours and synaesthetes select a colour for each grapheme (letters A-Z and numbers 0-9) three times in a randomised order. To measure consistency the distance of colours reported on each trial for each grapheme is calculated (in red, green and blue [RGB] colour space) and then averaged to create a distance score across all graphemes. This score is then used to determine if the person is a synaesthete or not, because a distance score below 1 is considered small enough to indicate genuine synaesthesia (i.e., small distance between the colours in test and retest means high consistency). Rothen, Seth, Witzel, and Ward (2013) later suggested that adjusting the cut off for synaesthetes to 1.43 may better distinguish between synaesthetes and controls, and is more in line with prevalence estimates (Carmichael et al., 2015). This gold standard of consistency testing remains the main way researchers define synaesthesia. However, the consistency test most commonly used is only applicable to adults, and for this reason adult synaesthesia has been the main focus in the literature. Here I use tests that apply these same principles in children. In other words, our child synaesthetes have been verified by the MULTISENSE project (Simner, Rinaldi, et al., 2019) using a test which elicits colours

for graphemes via a child-friendly colour-palette, and then set a threshold suitable to identify synaesthesia in children (see *identifying synaesthesia in children* below).

The second synaesthesia variant we focus on is OLP synaesthesia, and researchers have also investigated consistency in this type, where sequences such as letters and numbers are experienced as having personifications (i.e., personalities and/or genders). As with grapheme-colour synaesthesia, OLP synaesthetes' associations are consistent over time. Simner and Holenstein (2007) measured consistency for an adult with OLP by asking her to write down her personality associations for each grapheme. They found that, as with grapheme-colour synaesthesia, the OLP synaesthete was far more consistent over a retest interval of two years than controls were across 3 weeks. Amin et al. (2011) twice asked a group of 11 OLP synaesthetes to provide gender and/or personality associations for each of the graphemes they had associations for, along with a group of 11 controls who were asked to invent associations. Synaesthetes' consistency was measured over a period of a month or longer, whereas controls' consistency was measured between a period of 24 hours to a week. Again, synaesthetes were significantly more consistent than controls. Finally, Simner, Gartner and Taylor (2011) conducted a similar study with five OLP synaesthetes and corroborated these findings. The synaesthetes were retested after at least seven months and the controls after three weeks. Synaesthetes were significantly more consistent than non-synaesthetes. Thus overall people with OLP synaesthesia show the same pattern of consistency as those with grapheme-colour synaesthesia. As with grapheme-colour synaesthesia then, the test of genuineness used by the MULTISENSE project for our OLP synaesthetes (Simner, Alvarez, Rinaldi, Smees, & Carmichael, 2019) again identifies synaesthetes as those more consistent than their peers. To achieve this, the test required children to choose a specific personification for each grapheme. In adults (Hughes, Ipser, & Simner, 2019) this is done by eliciting each synaesthetic personality in depth using personality questionnaires (e.g., for each letter: how outgoing is it? How trusting? How thorough? How imaginative? Etc.) In children, the MULTISENSE test presented personifications in a simplified way. Hence, for each grapheme, children choose one of six different personifications represented by line-drawn faces (i.e., friendly female, unfriendly female, neutral female, friendly male, unfriendly male, neutral male). When children were asked to repeat this task some time later (see Chapter 3), true synaesthetes were identified as those who chose largely consistent personifications over time, while non-synaesthetes are inconsistent.

In this brief review, I have shown that consistency is the key principle that distinguishes synaesthetes from their non-synaesthetic peers. For this reason consistency is the main feature that has been used to verify that self-declared synaesthetes are truly synaesthetes. Here the fact that that we can apply these consistency principles to identifying child synaesthetes is taken advantage of, and I return to this point below after a short review of literacy and numeracy acquisition, as well as of the existing research on synaesthesia in children.

Letter and Number acquisition

The current thesis focuses on two types of synaesthesia in children, both of which are induced by grapheme information. Synaesthetes who have grapheme inducers may hold associations with either numbers only, letters only, or both letters and numbers. Those triggered by letters are dependent on literacy acquisition because children must logically learn the alphabet before associations can develop with letters. Therefore, it is appropriate to consider how the alphabet is acquired in order to identify the foundations of how synaesthesia itself is likely to be acquired.

In this thesis, we are testing children between the ages of 6-10 years. Worden and Boettcher (1990) examined alphabet knowledge in children just prior to this age-range, from 2.5 years to 7.5 years old. Whilst their youngest children could only recite on average up to five letters, by age 5 years children were able to recite almost all letters accurately. That said, there were some influence on the way letters were learned. Younger children were able to learn uppercase letters more quickly than lowercase letters, which are acquired later at ages 6-7 years. As well as naming, children were more easily able to print uppercase letters. This suggests that grapheme-colour synaesthetes may have coloured uppercase letters before they have lowercase letters, and for this reason, the MULTISENSE project used uppercase letters when identifying synaesthetes (whom we tested in Chapters 3 and 5).

Like with letters, the shapes of numbers are acquired during childhood, and therefore the trajectory of number acquisition is important to inform the age period in which synaesthesia is likely to develop (and I focus on numeracy skills in Chapters 4 and 5).

The acquisition of number words (e.g., “one”, “two”) involves, for instance, understanding that the word “five” corresponds to a specific quantity. This is a difficult task for a child, because most words refer to a property or an object in the environment rather than a set size. Yet children are already able to use number words by age 2-3 years, systematically understanding that they refer to a quantity (Hurewitz, Papafragou, Gleitman, & Gelman, 2006). Kuperman, Stadthagen-Gonzalez, & Brysbaert (2012) looked at the age of acquisition of 30,000 English words, including number words. From this data we can approximate that on average, children learn most number words between the ages of 3 and 5 years. In this thesis, I test children age 6 to 10 years and therefore I can be reasonably confident that children in our targeted age range would be able to complete our numerical cognition tests, and that child number synaesthetes are likely to have developed fixed (i.e. consistent) associations for some number graphemes. Further basis for this assumption is given in the section below, which describes the development of synaesthesia in children.

Synaesthesia in children

When synaesthetes are asked, they typically report that they have had their associations for as long as they can remember (Simner & Holenstein, 2007). Yet very little research has investigated synaesthesia in children. Here I summarise key studies and theories to explain the development of synaesthesia in children, to allow our research to be better contextualised.

The *Neonatal Synaesthesia Hypothesis* is a theory originally proposed by Maurer (1993) about how synaesthesia comes to develop in certain people. More specifically, the theory states that synaesthesia is experienced by all people as babies, but this experience is lost by most. The theory suggests that this natural state of synaesthesia in babies may arise due to functional hyper-connectivity in the brains of neonates. Importantly, there is evidence that both babies and adult synaesthetes have hyper-connected brains (Maurer, Gibson, & Spector, 2013; Rouw et al., 2011). In most people these connections are pruned away during the natural pruning process of cell death which takes place during normal development (Luo & O’Leary, 2005). In synaesthetes, however, the Neonatal Synaesthesia Hypothesis posits that these connections are not completely pruned away. There are a number of anecdotal reports in the literature of people who have had

synaesthesia during childhood and since lost their associations, which support this hypothesis (see Meier, Rothen, & Walter, 2014 for overview). There is also evidence which suggest that babies' sensory systems have much more cross-communication between different sensory modalities compared to adults. For example, Neville (1995) found that speech activates the visual cortex in infants and this response fades during the first three years. There is further evidence that in some adults this cross-connectivity remains; for example, grapheme-colour synaesthetes show increased connectivity (Rouw et al., 2011). Therefore, adult synaesthetes show similar patterns of hyper-connectivity as is found in all babies, which might provide support for the Neonatal Synaesthesia Hypothesis. Nevertheless, this hypothesis continues to be controversial (Deroy & Spence, 2013).

Aside from theories about how synaesthesia might emerge, there is clear evidence that it can be behaviourally detected by the age of 6 years. There are two studies of note, which show grapheme-colour synaesthesia within a childhood population, and the way in which it develops over time. Simner, Harrold, Creed, Monroe, and Foulkes (2009) identified child synaesthetes aged between 6-7 years old. They did this by asking children to give the "best colour" for each of the 26 letters of the alphabet (A-Z) and the numbers 0-9. Children chose from a palette of 13 colours, and were then given a surprise re-test after a 10-second pause. Children who showed signs that they may be synaesthetes (i.e. they were highly consistent between the initial test and surprise re-test) were revisited one year later -- and again when they were 10-11 years old (Simner & Bain, 2013). The number of stable synaesthetic associations grew over the time as the child aged. When the children were aged 6-7 years, approximately one third of their graphemes had consistent fixed colours (34%). By one year later when the children were 7-8 years old, almost half of the children's graphemes were fixed with consistent colours (48%). In the final visit, at age 10-11 years, children had fixed graphemes for almost three quarters of the graphemes (71%). In other words, over time, more of the colours for graphemes stabilised. However, by age 7-8 years, over half of the alphabet still did not have consistent colours, whereas adult synaesthetes typically have colours for close to 100% of their letters. This is important because in Chapters 2 and 5 we will see that the way in which child synaesthetes are identified (using consistency over time, see above) must take into account the fact that children's consistency grows as they age. Hence, a 6-year-old synaesthete may show similar consistency to a 10-year-old non-synaesthete, and so our

tests for synaesthetes from the MULTISENSE project take into account this development when identifying synaesthetes (Simner, Alvarez, Rinaldi, et al., 2019; Simner, Rinaldi, et al., 2019). In summary, the study by Simner et al. (2009) also paved a clear foundation in testing child synaesthetes. I carry forward the idea that synaesthesia develops over the course of childhood when using consistency testing with children.

The second longitudinal study of note is by Spector and Maurer (unpublished, reported in Maurer et al., 2013) who studied three pre-school aged children whose mothers have grapheme-colour synaesthesia. They were recruited on the assumption that they, too, were likely to develop synaesthesia (see Maurer et al., 2013). Children were given 96 crayons and told to use the most appropriate to colour in the 26 letters of the alphabet, the digits 0-9, and four basic shapes. One grapheme or shape was given per day to colour. Once all the stimuli were coloured, the cycle repeated. Children coloured the material over 3 or 6 cycles, and consistency between cycles was compared. These three children were far more consistent than age-matched non-synaesthetic controls, but significantly less consistent than adult synaesthetes, and consistency for later cycles was much higher than for the first cycles. These findings suggest, again, that synaesthetic colours stabilise over childhood, perhaps as children get more familiar with graphemes. However, Simner and Bain (2013) suggested the consistency of these very young synaesthetes (75% of letters were consistently coloured) was much higher than they would have predicted (less than 30%). Simner and Bain (2013) suggested that either the children recruited for this study had particularly strong synaesthetic experiences, or that it was an effect of the task itself which is very repetitive colouring letters daily for an extended period. This study also highlights the importance of the methodology used to identify child synaesthetes. Simner and Bain have argued that any child-synaesthete who is referred for recruitment by their parents may not represent synaesthetic children more broadly. This is because these parents are willing to reach out to researchers for science studies (and their family might reasonably be considered different from *average* families; and we discuss this in more detail in Chapter 3). This thesis used a random sampling method for this reason. Our child synaesthetes were identified from the MULTISENSE screening program, which screened all children within Years 2-5 of primary school from 22 schools (with only 1% withdrawal). Synaesthetes were therefore identified from a random sample, and also with a child-oriented diagnostic test which I describe below.

Identifying synaesthesia in children: The MULTISENSE project

Thus far I have summarised the current research on childhood synaesthesia in terms of consistency testing, the development of letters and numbers, and synaesthesia in children. Here I revisit this question of how to test objectively for synaesthesia in the context of testing children by considering the MULTISENSE project. I summarise the wider context of the grant that funded the current thesis along with tests we have developed in the project (e.g., Simner, Alvarez, et al., 2019; Simner, Rinaldi, et al., 2019), which I will use in the thesis to diagnose child synaesthetes.

This thesis forms part of the MULTISENSE project, a research program funded by the European Research Council to investigate synaesthesia during childhood. Within this, the MULTISENSE team (including the author of this thesis) conducted a large longitudinal study on childhood synaesthesia with three key aims. Firstly, we created and validated tests to identify child grapheme-colour synaesthetes and child OLP synaesthetes. Secondly, we investigated differences between synaesthetes and non-synaesthetes in various aspects of their cognition, personality, and wellbeing. And lastly, we created resources for parents and teachers to aid in the understanding of synaesthesia. This thesis focuses in the second strand, investigating what child synaesthetes are like in two areas: cognition (here, specifically, numerical cognition) and personality. However, since I was also been heavily involved in the first strand, creating and validating tests to identify child synaesthetes, I briefly describe this below where it is relevant to this thesis.

I noted above that diagnostics for synaesthesia rely on consistency tests (i.e., they identify synaesthetes by looking for the marker of consistency-over-time, in the colours of letters, personalities of numbers, and so on). I additionally noted that consistency tests such as those by Eagleman et al. (2007; for grapheme-colour synaesthesia) and Hughes et al. (2019; for OLP synaesthesia) have been designed for, and validated on, adult populations. As a result, both the grapheme-colour and OLP tests would be difficult for children to complete. The Eagleman grapheme-colour test has a small interface and a complex colour palette that is difficult for children to navigate, and the OLP task requires a complex understanding of personality (and personality vocabulary; e.g., how “thorough” is each grapheme?) beyond the capabilities of a young child. Moreover, both tasks are time consuming. The MULTISENSE project therefore aimed to design tasks appropriate for children.

The central problem in designing a test for children is finding one that can recognise synaesthetes as being different to non-synaesthetes. As child-synaesthetes only have a small percentage of their graphemes fixed at a young age, any test designed with children in mind would need to account for this. For example, the adult test requires that the synaesthete has most if not all graphemes consistently coloured. But a child test needs to account for the fact that children are only going to be consistent on the graphemes that they have fixed at that age (without the experimenter knowing what these graphemes would be). Additionally, whilst the earliest test for children (Simner et al., 2009) was sensitive enough to find certain synaesthetes, it did not contain a very sophisticated colour palette (there were only 13 colours). This means that it may not have had the sophistication to detect *all* developing synaesthetes. For the grapheme-colour task therefore the MULTISENSE project aimed to design a test as sophisticated as adult tests (e.g., large array of colours), but with an intuitive age-appropriate interface, and a shorter length of test (appropriate for the attention of a child). Additionally, the MULTISENSE team aimed to create a test that would take the age of the child into account, and give adjusted consistency scores that account for age. Although the MULTISENSE diagnostic of synaesthesia in children is to be reported elsewhere in full (Simner, Alvarez, Rinaldi, et al., 2019; Simner, Rinaldi, et al., 2019), in Chapter 3 we describe this test in enough detail for the purposes of this thesis. Specifically, where Simner et al. present in-depth details of the testing interface (e.g., motivations for design choices) and of scoring protocols (e.g., a variety of ways to compute scores for synaesthetes, and ways these might suggest synaesthesia at different stages in testing), my own studies describe the test in as much detail as the reader needs to be certain we have adequately identified synaesthetes.

The MULTISENSE project had similar aims for OLP synaesthetes. This was to create the first diagnostic of OLP synaesthesia for children using similar consistency measures. In order to do this, it was necessary to simplify the concept of personality to valence (positive, neutral, and negative) which we expressed by line-drawings of faces that were normed to appear either friendly, neutral, or unfriendly. Children were required to match one face to each grapheme. This task also recognised that it is possible to be an OLP synaesthete in three ways: to have gender associations only, personality associations only, or both. Again, the test gave a consistency score that accounts for age (i.e., identifies a 6-year-old and 10-year-old synaesthete differently, according to the consistency expected

at each age). In Chapter 3 we again provide sufficient details of how the MULTISENSE test achieves this.

In summary, in the MULTISENSE longitudinal project, two tests were created to diagnose childhood synaesthesia. Both a grapheme-colour test and an OLP test were created, and we describe both tests in this thesis, in Chapters 3 and 5. We also tested synaesthetes and non-synaesthetic controls on a wide range of cognitive and academic tests, including tests related to literacy, creativity, visuospatial awareness, working memory, receptive vocabulary, wellbeing, numerical cognition, and personality. In this thesis I focus on my work in just two of these domains: numerical cognition and personality. Importantly, all of these tests were carried out in parallel with the diagnostics for synaesthesia. This means that none of the experimenters knew which of their participants were synaesthetes. The testing was therefore ‘blind’ in this regard, and experimenters could therefore not influence the outcomes of the cognitive and personality testing described in this thesis.

Environmental synaesthesia and unusual associations in non-synaesthetes

In Chapter 4 we look at a widely used colour-number tool which is used in primary schools to help children with mathematics. This tool, pairs colours to numbers (e.g., number 5 is red) and a small number of children learn these associations, giving them synaesthetic-like experiences. At the same time, true synaesthetes can also learn their colours from the environment, in certain ways. In this section I therefore give an overview of environmental influences in synaesthesia, and how true synaesthetes differ from the non-synaesthetes tested in Chapter 4.

Whilst synaesthesia has a genetic component (see Asher et al., 2009), environmental ties have been noted. A study by Mankin and Simner (2017) found that some of the letter-colour trends found across synaesthetes (e.g., A tends to be red above all other colours) can be tied to common associations taught to children during literacy acquisition (e.g., A is for apple; apples are red). Another key example of environmental influence is a case of an adult non-synaesthete who learned colour-number associations due to cross-stitching correspondences (Elias, Saucier, Hardie, & Sarty, 2003). This individual also showed similarities to a synaesthete in a synaesthetic Stroop task (i.e., both the synaesthete and the cross-stitch associator were slower to respond to numbers presented in colours

incongruent with their associations). They differed, however, in fMRI activation when shown achromatic digits; the synaesthete showed additional activation in the left dorsal visual stream, not seen in the cross-stitch colour associator. This suggests that our children in Chapter 4, who have internalised the colours of their educational maths tool, will show behavioural similarities to synaesthetes, even if underlying neural architecture is different (see Chapter 4).

Complicating the picture somewhat is that true synaesthetes can, in fact, take their colours from environmental sources. Evidence comes from a group of synaesthetes known as “refrigerator magnet synaesthetes,” so called because their associations can be traced back to a popular fridge magnet set that was produced between 1971 and 1990. Of the data from over 6500 synaesthetes, 400 could be traced back to this schema. During the period that the refrigerator magnets were produced, 9.1% of the synaesthetes born were magnet synaesthetes (Witthoft, Winawer, & Eagleman, 2015). Given these environmental links, several studies have attempted to train adult non-synaesthetes to have coloured letter and number associations with mixed results (Kusnir & Thut, 2012; Macleod & Dunbar, 1988; Meier & Rothen, 2009; Rothen, Wantz, & Meier, 2011). Bor, Rothen, Schwartzman, Clayton and Seth (2014) trained non-synaesthetes over a longer course of nine weeks. After this time, 9 out of 14 of the participants reported phenomenological experiences to the trained letters, such as seeing the colours in the real world, although these experiences tended to die away after the study ended. But such studies raise the question of how to distinguish between a true synaesthete, and a non-synaesthete who has internalised environmental colours. True synaesthetes will have internalised those colours spontaneously (without teaching being necessary) and from an early age, will experience them automatically without effort, and have lifelong associations, which are consistent over time. They will also (given the findings of Elias et al., 2003) be relying on different neurological architecture. Nonetheless, these studies clearly show that whilst there may be a genetic component to synaesthesia, there is also a role for learning during childhood, and this is particularly pertinent in Chapter 4, where we look at a number-colour tool that is being actively taught in schools.

Overview of experimental chapters

Throughout the thesis, I explore what makes child synaesthetes different, aside from their synaesthesia. Research into the development of synaesthesia in child populations has been few and far between, due to the difficulties accessing the population and the lack of tools available to identify synaesthesia in childhood populations. I use rigorous testing methodologies with purpose-built child tests. I present data from 80 randomly sampled synaesthetes, the largest cohort of child synaesthetes the field has seen to date. All the chapters in the thesis stem from one longitudinal project with regards to data collection and testing. Figure 1 below shows how the different tests used throughout the thesis relate to the different stages of testing.

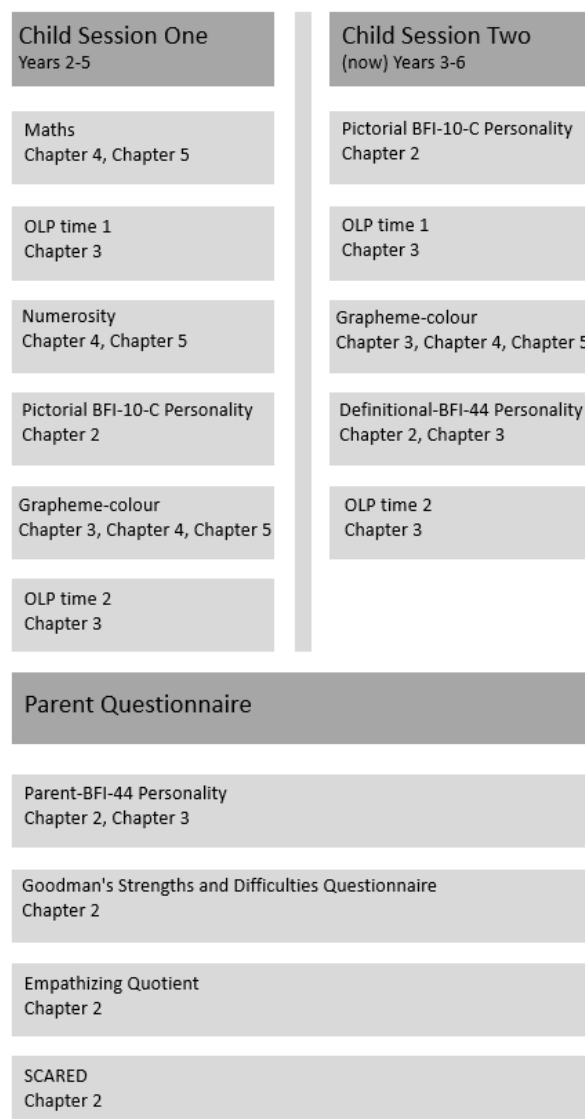


Figure 1: MULTISENSE data collection detailing each test used in the thesis broken down by Session and Chapter. As the figure indicates, parent questionnaire responses spanned both rounds of testing.

In Chapter 2, I will start by validating a tool that has been designed for the purposes of this thesis. Throughout the thesis, I encountered a lack of appropriate tools available for testing children aged 6-10 years old. This was most prominent in the personality field, where instruments designed for children have been typically aimed at those in Secondary school aged 12 and above (e.g., Big Five Questionnaire for children; Muris, Meesters, & Dideren, 2005). The few resources designed for younger children were time-consuming and inappropriate for testing many thousands of children, as was necessary here. For example, the Berkeley puppet interview (Ablow & Measelle, 1993) can only test children one by one (not in parallel) and is relatively long. Therefore, in Chapter 2, I created a novel test for personality designed with young children in mind. I also took two existing Big Five Inventory questionnaires and validated both for our sample populations. One was a test for adolescents, which I adapted for younger children. The other was a test for parents which had never been validated, so I validated that also. In Chapter 3, I used my validated personality questionnaires from Chapter 2 to look at differences in personality between grapheme-colour synaesthetes, OLP synaesthetes, and non-synaesthetic controls. In this chapter, we noted that there is emerging evidence to suggest that adult synaesthetes have a particular personality profile (Banissy et al., 2013; Chun & Hupé, 2016; Rouw & Scholte, 2016). I therefore investigated personality in a child sample for the first time.

In Chapters 4 and 5, we depart from personality and ask whether there are differences between synaesthetes and non-synaesthetes in cognition. Numeracy is an area that allowed us to additionally explore synaesthetic-like associations in non-synaesthetic children. In Chapter 4, we examined a tool used by many primary schools throughout the UK called *Numicon* (Oxford University Press, 2018). This tool uses physical objects (coloured plastic shapes) to represent each of the numbers between 1 and 10 and is a tool used to aid in teaching. Importantly, it holds specific colour-number pairings (e.g., the shape for 5 is red). We explored whether any children had internalized these colours by asking children to give colours for numbers, and identifying children who matched the Numicon schema at higher than chance levels. If they did match to Numicon, we asked whether having these colour associations improved their numerical cognition compared to controls who did not use a strategy. We considered two aspects of numerical cognition: mathematics and numerosity. Numerosity is our intuitive number sense that allows us to estimate the number of objects in an array when we do not have enough time to count

them (Dehaene, 2001). We examined whether having internalised Numicon colours (rather than simply having been exposed to Numicon) leads to benefits in either aspect of numerical cognition.

In Chapter 5 we then turn to numeracy skills in synaesthetes themselves. Here we tested grapheme-colour synaesthetes and explored whether synaesthetes showed any differences from non-synaesthetes in the same tests of numerical cognition used in Chapter 4. In order to further investigate a model we proposed to explain findings in Chapter 4, we specifically looked separately at grapheme-colour synaesthetes with coloured numbers versus coloured letters. In each case, we compare whether synaesthetes perform any differently compared to controls.

Summary

In this thesis, I will explore two key domains which have been tied to differences in adult synaesthetes, and show that randomly sampled child synaesthetes show particular personality profiles and differences in numeracy skills. In the following experimental chapters, I will validate personality questionnaires, test grapheme-colour and OLP synaesthetes identified using random sampling, and use my newly validated measures to show differences between synaesthetes and non-synaesthetes. I additionally identify non-synaesthetic children that have internalised colours from a number-colour educational tool and examine differences in numerical cognition in these non-synaesthetes and in our randomly sampled synaesthetes. Each of these experimental chapters are presented in a *paper style* format, and three of the four chapters are currently under peer-review or have been accepted for publication. I present a short summary at the beginning of each chapter to tie together the experimental chapters that make up this thesis.

Chapter 2

Big Five Personality Instruments for Parents and Children 6+ years: The Pictorial BFI-10-C; the Definitional BFI-44-C, and the BFI-44- Parent.

Chapter Summary

In this thesis I am interested in exploring child synaesthetes and their differences. The first focus in the thesis is differences in personality. There has been some evidence to support personality differences in adults, and I aim to extend and clarify these findings in children. Before I turn to exploring differences between synaesthetes and non-synaesthetes however, I first validate three personality instruments in this chapter for use with children aged 6-10 years. I then use these measures to explore differences in personality in synaesthesia in Chapter 3. In the current chapter, I take two existing measures of personality; one parent-rated questionnaire that has not yet been validated, and one child-rated questionnaire for adolescents that I adapt to suit a younger demographic. I additionally create and validate a new questionnaire designed specifically with younger children in mind that pictorially presents personality. This chapter has been submitted for publication as Rinaldi, L.J., Smees, R., Carmichael, D.A., & Simner, J (2019) *Big Five Personality Instruments for Parents and Children 6+ years: The Pictorial BFI-10-C; the Definitional BFI-44-c, and the BFI-44-parent*. Manuscript submitted for publication. Note that where additional models were included in a supplementary information in our article submission, here they have been instead provided at the end of the chapter.

Abstract

In this paper we created and/or validated three *Big Five Inventory* personality questionnaires for children in middle childhood. We first developed a novel 10-item pictorial self-report questionnaire (*The Pictorial-BFI-10-Child*) and validated this on 3349 children age 6-10 years. Next, we adapted an existing questionnaire, the BFI-44-*Adolescent*, to make it suitable to an audience as young as 8 years (and we named this revised instrument the *Definitional BFI-44-Child*). We validated this revised questionnaire on 846 children age 8-11 years. Finally, we validated an existing (unchanged) parent questionnaire (*BFI-44-Parent*) on 550 parents of children aged 6-10 years. Results show that all three measures validated well: our *Pictorial-BFI-10-Child* had good test-retest reliability, good concurrent validity, and expected levels of convergent validity. The *Definitional BFI-44-Child* had good internal reliability, and expected concurrent and convergent validity, and the *BFI-44-Parent* had excellent internal reliability, and expected convergent validity. We conclude from our analyses that children aged 8-11 years can be reliably assessed using all three tools. We conclude too that yet-younger children age 6-7 years can provide personality self-report (with the *BFI-10-C-Pictorial*) but are reliable indicators only of their own *Agreeableness* and *Neuroticism*. These younger children are therefore better evaluated more fully in conjunction with a parent-led questionnaire (*BFI-44-Parent*).

Introduction

One elegant way to measure personality is to consider it as having component parts, or *factors*. Tupes and Christal (1961) defined five factors of personality which were later refined (e.g., by Goldberg, 1990; McCrae & Costa, 1987) into five categories widely known today as: *Openness to Experience*, *Conscientiousness*, *Extraversion*, *Agreeableness* and *Neuroticism*. *Openness* is typically associated with intellect and creativity (Caspi, Roberts, & Shiner, 2005), *Conscientiousness* is associated with hard work and carefulness (McCrae & Costa, 1987), *Extraversion* is associated with being outgoing and sociable (McCrae & Costa, 1987), *Agreeableness* is associated with trust and sympathy (McCrae & Costa, 1987) and *Neuroticism* is associated with worry and emotional instability (McCrae & Costa, 1987). These ‘Big Five’ personality traits can be measured in adults using questionnaires which vary in length from very brief (e.g., 5-10 item inventories: Gosling, Rentfrow, & Swann, 2003; Rammstedt & John, 2007) to relatively long (e.g., NEO-FFI; Costa & McCrae, 1992). But personality traits can also be measured across the lifespan (Roberts & DelVecchio, 2000) and within the last ten years in particular, similar personality research has been extended to children (Mackiewicz & Cieciuch, 2016; Markey, Markey, & Tinsley, 2004; Markey, Markey, Tinsley, & Ericksen, 2002; Measelle, John, Ablow, Cowan, & Cowan, 2005; Muris et al., 2005). Here we contribute to this literature with three novel and/or newly-validated personality tests for children age 6+ years.

A number of questionnaires exist to measure the Big Five traits in children, such as the *Big Five Inventory* (BFI-44-A; John, Donahue, & Kentle, 1991; John et al., 2008; John & Srivastava, 1999), the *Big Five Questionnaire* (BFQ-C; Barbaranelli et al., 2003), the *Inventory of Child Individual Differences* (ICID; Halverson et al., 2003), the *California Child Q Set* (John, Caspi, Robins, Moffitt, & Stouthamer-Loeber, 1994), and the *Hierarchical Personality Inventory for Children* (HiPiC; Mervielde & De Fruyt, 1999). However, these questionnaires are typically filled out by a close adult such as a parent or teacher (e.g., California Child Q set and the HiPiC). There are very few self-report questionnaires for completion by children and those that exist are relatively long (44 to 65 items at least) (Barbaranelli et al., 2003; John et al., 2008) and typically aimed at older children 10+ years (e.g., Barbaranelli et al., 2003; Soto et al., 2008). Moreover, there is only moderate agreement between child-reported and mother-rated personality, at least in children aged approximately 10 years (Markey, Markey, Tinsley & Ericksen, 2002).

Similarly, McCrae and Costa (1987) found reasonably low levels of agreement between personality assessments from self-report and peer-report. This suggests that children's self-rated personality may hold additional information, or that children may have a different view point, compared to ratings of them by others. At the very least, it may be beneficial to use a child's own self-report in conjunction with adult-rated personality, to get a more comprehensive assessment of personality in children.

Other methods of gathering personality information directly from children are interview techniques, such as the *Berkeley Puppet Interview* (Ablow & Measelle, 1993), but although these work well for younger children (Measelle et al., 2005) they require one-to-one testing so are time consuming. In the current study we introduce two novel ways to test young children (age 6+ or 8+ years) by their own self-report, and we additionally validate an existing parent questionnaire (for children 6+ years) which to our knowledge has not been purposefully validated previously. Our first self-report measure was designed for a 6+ year old audience, and is based on the adult *Big Five Inventory 10-item questionnaire* (*BFI-10*; Rammstedt & John, 2007) but with items presented for children in pictorial form. We devised this questionnaire then tested it on children aged 6-10 years. Our second self-report measure is a new adaptation of an existing self-report questionnaire, the *BFI-44-Adolescent questionnaire* (*BFI-44-A*; John, Donahue, & Kentle, 1991; John et al., 2008; John & Srivastava, 1999). Here we added definitions to words in the questionnaire in order to make it more easily interpretable for younger children, and we tested it on children aged 8-10 years. Finally, we took an unchanged existing questionnaire for parents, the *BFI-44-parent* (produced by the Berkeley Personality lab; see John, Donahue, & Kentle, 1991; John et al., 2008; John & Srivastava, 1999), which to our knowledge has not been purposefully validated as a parent questionnaire for children. Below we start with a brief overview of the relevant personality literature, including a consideration of the challenges in testing personality during middle childhood.

There are a number of considerations for the design of a self-report personality questionnaire for children. The first is the question of whether personality is stable at this age, and whether children have the level of introspection to reliably report it. Understanding this issue is one aim of the current study. Compared to personality in adults, (which is relatively stable despite some changes; e.g., older people increase in *Conscientiousness* and decrease in *Openness*) childhood personality may be less stable

(Caspi et al., 2005). It has nonetheless shown moderate consistency when personality is compared over time (Caspi et al., 2005) and is stable enough for measurement. For example, using the Berkeley Puppet Interview, Measelle et al. (2005) found that children's self-reported personality was as consistent as a college sample. These findings additionally suggest that children can reliably inform about their personality from as young as five years of age, at least in time-intensive interview techniques.

A second consideration when developing a child questionnaire is length. In children – and indeed in adults – there has been a growing need for short questionnaires, which can gain a broad sense of personality in research situations where time is limited (Rammstedt & John, 2007). For example, the TIPI and the BFI-10 (Gosling et al., 2003; Rammstedt & John, 2007) are ten-item adult personality questionnaires for any research situation which would preclude longer testing (e.g., when testing large numbers or where time and/or attention is limited). Hence, although longer questionnaires often have greater validity, researchers have nonetheless recognised the need for a short measure (e.g., Rammstedt & John, 2007). Crede, Harms, Niehorster, and Gaye-Valentine (2012) showed that despite limitations (e.g., increases in type 1 and 2 error rates), shorter questionnaires should still be considered for populations who have an increased likelihood of disinterest or fatigue. And nowhere are these constraints more obvious than when testing young children. Children fatigue more easily than adults, struggle more with reading, and take more time to answer each question given underdeveloped cognitive skills. This gives a particularly pressing need for child questionnaires that are short.

A third consideration is how to control for acquiescence. An acquiescence bias is shown when participants consistently say yes (or consistently say no) to items which are logical opposites (e.g., “is talkative” and “tends to be quiet”). Studies have shown that this response style is more common amongst low-educated samples, such as child samples (Meisenberg & Williams, 2008; Rammstedt & Farmer, 2013). In personality research in particular, several studies have shown that it is important to adjust for acquiescence (Danner, Aichholzer, & Rammstedt, 2015; Rammstedt & Farmer, 2013; Soto & John, 2009) and particularly as the sample gets younger (Soto et al., 2008). In the current study we test even younger children than is typical, and adjust for acquiescent responding accordingly.

A fourth challenge for creating a testing instrument for middle childhood is whether to measure “Big Five” personality or other personality-related indices (e.g., *temperament*; a biological predisposition towards certain behaviours; Goldsmith et al., 1987). In the current study we focus on the Big Five Inventory (John et al., 2008) for a number of reasons. Firstly, studies show that many dimensions are similar between temperament and personality, and may in any case overlap (Goldsmith et al., 1987). For instance, Caspi and Shiner (2006) suggested that *Neuroticism* is consistent with the temperament trait of *Negative Affectivity*, *Extraversion* is consistent with *Surgency* and *Conscientiousness* is consistent with *Effortful Control*. Second, a Big Five questionnaire already exists which has the potential to be adapted for a middle childhood audience: the BFI-44-A. Although designed for adolescents, the adult equivalent of this test, the BFI-44, has already been validated in children as young as ten years (Soto et al., 2008), and the BFI-44-A is much shorter than many personality questionnaires, which sometimes range from 60 to 100 items (e.g., Barbaranelli et al., 2003; John et al., 1994). This questionnaire might therefore be adaptable for a yet-younger audience (here, 8+ years).

The final challenge when developing a self-report questionnaire for children in middle childhood is how to convey the relatively abstract concepts of personality to children via a more concrete mode of presentation. Since as far back as Piaget (1965) at least, we have known that children of this age group may need additional help to understand abstract ideas. Here we accompany our personality testing (in one instrument at least) with pictures representing the personality traits under investigation, not only to reduce task demands of reading but also to present a more concrete representation of abstract ideas. Below we summarise the three personality instruments we will test in the current study.

Three New and/ or Newly-validated Instruments

BFI-44 Parent. This 44-item personality test is available from the Berkeley Personality lab (<https://www.ocf.berkeley.edu/~johnlab/measures.htm>), and is designed for parents to assess the personality of their children. Here we will validate this test on the parents of children as young as six years. To our knowledge, this questionnaire has not been

validated on parents before, although a related instrument has been well-validated by self-report from adults (BFI-44; John et al., 2008)².

Definitional BFI-44-Child. We have created this questionnaire from an existing measure, the BFI-44-Adolescent, which is available from the Berkeley Personality lab (<https://www.ocf.berkeley.edu/~johnlab/measures.htm>), (BFI-44-A; John et al., 1991, 2008; John & Srivastava, 1999; Soto et al., 2008). Here we altered this questionnaire to provide definitions of late-acquired words, in order to make the instrument suitable for testing younger children. For instance, for the item “I see myself as someone who generates a lot of enthusiasm,” we added a definition for the final word (“This means getting excited about things”). We will validate on groups of children between the ages of 8 and 11 years.

Pictorial BFI-10-Child. This is our own novel self-report measure which we will validate on groups of children between the ages of 6 and 10 years. Prior to our study there had been just one existing pictorial personality questionnaire for children: the *Pictorial Personality Traits Questionnaire* for children (PPTQ; Mackiewicz & Cieciuch, 2016). This Polish-language self-report questionnaire was validated by those authors on a group of 501 Polish children aged 6-12 years (in two groups; 6-9 years, and 10-12 years). Importantly, their youngest age group had a mean age of 9.25 years, so was highly skewed to older children. Our own English language questionnaire will be validated on English-speaking children using a younger cohort (see Methods; Study 2). Mackiewicz and Cieciuch (2016) found their Polish instrument had problems with internal reliability, and our own measure attempts to improve on several design features. Like our own questionnaire, the PPTQ represents pictorial personality questions where text is accompanied by two pictures per item (e.g., *Extraversion*’s “I usually play with others/alone” comprises a left-hand picture showing the protagonist playing alone and a right-hand picture showing the protagonist playing with other children). Importantly, their pictured events differed on extraneous elements which could influence responding in

² The relationship between BFI questionnaires can be confusing so is clarified here. The BFI-44 (John et al., 2008) is an adult self-report questionnaire whose questions begin “I see myself as someone who...” (e.g., “I see myself as someone who prefers work that is routine”). The BFI-44-Adolescent is the same questionnaire with minor changes for adolescents (e.g., “I see myself as someone who likes work that is the same every time (routine)”). The BFI-44-Parent is the same as the adolescent questionnaire but this time asking about the child (e.g., “I see my child as someone who likes work that is the same every time (routine)”). Here we take the BFI-Adolescent and expand it with definitions suitable for yet-younger children.

unwanted ways (e.g., the child playing alone is on a swing but the child playing with others is in a sandpit). This could lead young children to choose which activity they prefer rather than whether they like playing alone or not. In our own questionnaire, both pictures within each item were identical other than for the personality trait of interest. Second, like our own questionnaire, the PPTQ shows a child-protagonist designed to be visually gender-neutral. However, this was not pre-tested, and our own norming shows that 100% of our participants ($n = 10$) perceived their protagonist to be male. It is important to ensure that any differences found in personality traits across the sexes do not arise from confounds in self-identification, so our own protagonist was pre-tested to be gender neutral (see Methods; Study 2).

We also reduced the length of our questionnaire overall: the PPTQ has 15 items and our own is based on a prior 10-item questionnaire (BFI-10). Finally, we introduced a greater opportunity for detailed responding, even from younger children. The PPTQ allows children to respond on a three-point (children aged 6-9) or 5-point scale (children aged 10-12). In our own study we devised a way for children aged 6-10 years to respond relatively easily with an eight-point scale. For each item, children first choose between two pictures (“I prefer playing with other children/ on my own”) and then say *how much* the choice is like them (“Just a bit/ Sometimes/ Mostly/ Completely”). This allowed us to gain a greater range of responses from children whilst simplifying the response into a binary decision (i.e., first choosing a picture) followed by 4-point responding. With these considerations in mind we created our new measure the Pictorial BFI-10-C.

Study 1 seeks to validate the BFI-44 parent instrument, then Study 2 seeks to validate both the Definitional BFI-44-C and our new measure, the Pictorial BFI-10-C.

Study 1: Validating the Big Five Inventory-44-Parent

Methods.

Participants.

In total we sent our questionnaire to the parents of 3349 children aged between 6 to 10 years (mean age 7.92, SD 1.22). Of these children, 1639 were girls (mean age = 7.91, SD = 1.23) and 1707 were boys (mean age = 7.94, SD = 1.21). These children/parents were recruited from 22 local primary schools in East and West Sussex, UK. Our respondents

(i.e., our final participants) were the parents of 550 of these children (6-10 years; mean age = 8.37, SD = 1.14). Of these children, 263 were girls (mean age = 8.37, SD = 1.13) and 287 were boys (mean age = 8.37, SD = 1.15). Ethical permission for all studies reported here was granted by the local ethics board (Sussex University Science and Technology Research Committee reference ER/JCS41/5).

Materials and Procedure.

Our parent-questionnaire was sent out in October 2017, and then again (as a reminder) approximately 6-10 months later. Parents completed their questionnaire either in an online electronic form ($n = 401$) using the testing platform Qualtrics or in a pencil-and-paper version ($n = 149$). This decision was dictated by whether participating schools contacted their parent-body electronically via email or using paper-letters. All tests were presented identically electronically or on paper.

The BFI-44-Parent. The BFI-44-Parent is the parent version of the BFI-44-Adolescent (Big Five Inventory Adolescent questionnaire). The questionnaire contains 44 statements linked to Big Five personality traits (plus a two item “liking” scale which we omitted³). Each statement begins “I see my child as someone who...” There are ten items for *Openness*, nine each for *Agreeableness* and *Conscientiousness* and eight items each for *Extraversion* and *Neuroticism*. For example, Item 3 relates to *Conscientiousness* and states “I see my child as someone who does things carefully and completely.” Some statements are positively worded with respect to their trait and some are negatively worded (e.g., Item 6 relates negatively to *Extraversion*: “I see my child as someone who is reserved; keeps thoughts and feelings to self”). Parents are asked to indicate how much they equate their child with each statement, using a five-point Likert scale “Disagree strongly/ Disagree a little/ Neither agree nor disagree/ Agree a little/ Agree strongly”. The measure takes between 5-10 minutes to complete.

³ The full questionnaire is known as the BFI-46 and comes with a two-item liking scale, which asks how much children are liked by others. We omit this two-item scale here as we are interested in only in the Big Five personality traits.

In order to validate the BFI instruments we also asked parents to complete the following additional questionnaires. Each questionnaire will be used as a measure of convergent validity for our personality questionnaires.

Goodman's Strengths and Difficulties Questionnaire (SDQ). The SDQ (Goodman, 1997) is a 25-item questionnaire which assesses behaviour on five factors comprising *Conduct Problems*, *Hyperactivity*, *Emotional Symptoms*, *Peer Relationship Problems* and *Prosocial Behaviour*. Each item is presented as a statement and parents, rate on a three-point Likert scale ("Not true/ Somewhat true/ Certainly true"), based on their child's behaviour in the last six months. For example, Item 10 relates to hyperactivity and states "Constantly fidgeting or squirming". The questionnaire takes approximately five minutes to complete.

The Empathy Quotient (EQ). The EQ (Auyeung et al., 2009) is a 27-item questionnaire in which parents assess their children's levels of empathy. Questions are presented as statements such as "My child likes to look after other people." Parents respond on a four-point Likert scale "Definitely agree/ Slightly agree/ Slightly disagree/ Definitely disagree." The questionnaire takes approximately five minutes to complete.

The Screen for Childhood Anxiety Related Disorders (SCARED). The SCARED (Birmaher et al., 1999, 1997) is a 41-item anxiety screening questionnaire, used to identify symptoms related to *panic disorder*, *general anxiety disorder*, *school avoidance*, *social anxiety* or *separation anxiety*. Questions are presented as statements, which parents rate based on their child over the past three months. For example, Item 36 relates to school avoidance and states "My child is scared to go to school." Parents respond on a three-point Likert scale: "Not true or hardly ever true/ Somewhat true or sometimes true/ Very true or often true." The questionnaire takes approximately five to ten minutes to complete.

Results

Analyses below are based on 550 parent-responders unless otherwise stated (i.e., where parents omitted sections of our questionnaire, participant numbers are stated for each analysis).

Construct Validity. Construct validity relates to whether our items load onto an expected construct (personality in this case). Here we report exploratory factor analyses due to known issues with confirmatory factor analyses (CFA) in personality research (Hopwood & Donnellan, 2010). In particular, CFA can penalize large sample sizes and has limited capacity in accounting for the complex nature of personality (e.g., personality items for a particular factor may show cross-loadings to other factors, which is problematic for CFA).

Our exploratory factor analysis was a principal components analysis (PCA) on the 44 items using varimax orthogonal rotation. Sampling adequacy was “excellent” with a Kaiser-Meyer-Olkin (KMO) score of .90. We extracted five components based on our a-priori assumptions. The scree plot supported a five-components solution (see Figure 1) which accounted for 49.73% of the variance. Table 1 shows the components loadings after rotation. These component loadings support evidence of the Big Five factors of *Openness*, *Conscientiousness*, *Extraversion*, *Agreeableness* and *Neuroticism*, although one *Openness* item (“I see my child as someone who likes work that is the same every time (routine)”) loaded higher onto *Neuroticism* than onto *Openness*.

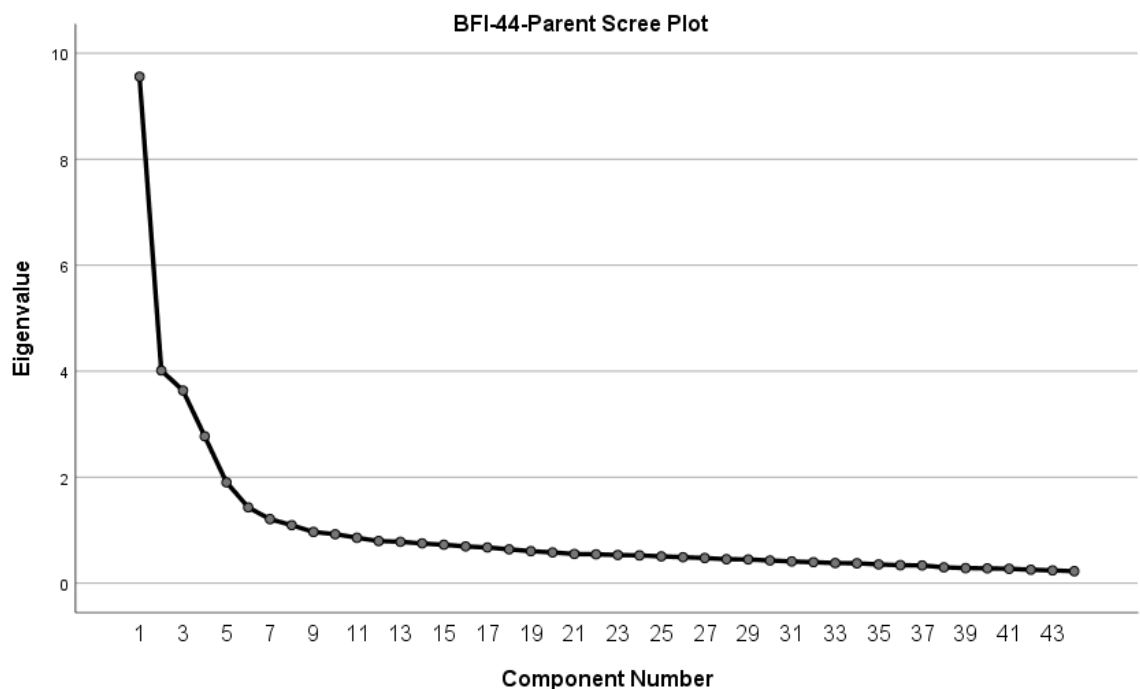


Figure 1: Scree plot suggesting a 5-component solution for data from the BFI-44-Parent questionnaire

Table 1.

Component loadings for unadjusted BFI-44-Parent scores. Table shows loadings over .10 with strongest component loadings in bold. Item numbers correspond to BFI-44-parent questionnaire; see Berkeley personality lab for questionnaire (John, 2009). For header, O=Openness; C=Conscientiousness; E=Extraversion; A=Agreeableness; N=Neuroticism.

Item		<i>O</i>	<i>C</i>	<i>E</i>	<i>A</i>	<i>N</i>
<i>Openness</i>						
35	Likes work that is the same every time (routine)	-.14			-.12	.27
25	Is creative and inventive	.79	.14			-.16
30	Likes artistic and creative experiences	.74				
5	Is original, comes up with new ideas	.72		.12		-.14
20	Has an active imagination	.72	-.11			
40	Likes to think and play with ideas	.70	.16	.14	.18	
44	Knows a lot about art, music, or books	.50	.11	.13		
41	Doesn't like artistic things (plays, music)	-.49				
15	Is clever, thinks a lot	.49	.36	.21	.12	
10	Is curious about many different things	.45	.15	.30	.23	
<i>Conscientiousness</i>						
3	Does things carefully and completely	.17	.74			
33	Does things efficiently (quickly and correctly)	.16	.71	.11		-.19
18	Tends to be disorganized	.17	-.69			.15
13	Is a reliable worker	.17	.68		.28	
28	Keeps working until things are done	.17	.68		.11	-.12
43	Is easily distracted; has trouble paying attention		-.63		-.11	.23
8	Can be somewhat careless	.14	-.57	.11	-.14	.19
38	Makes plans and sticks to them	.11	.56	.13		
23	Tends to be lazy	-.10	-.38		-.23	
<i>Extraversion</i>						
21	Tends to be quiet		.12	-.78		.20
36	Is outgoing, sociable	.15		.74	.23	-.18
1	Is talkative	.24		.72		
6	Reserved; keeps thoughts and feelings to self			-.63		.18
31	Is sometimes shy, inhibited			-.63		.36
26	Takes charge, has an assertive personality	.18	.21	.63	-.23	-.13
16	Generates a lot of enthusiasm	.41	.15	.57	.28	
11	Is full of energy	.27		.56		-.11
<i>Agreeableness</i>						
32	Is considerate and kind to almost everyone	.17	.20	.17	.74	
37	Is sometimes rude to others		-.12	.15	-.69	.22
7	Is helpful and unselfish with others	.13	.19	.22	.66	
17	Has a forgiving nature	.11		.14	.66	-.15
12	Starts quarrels with others			.13	-.65	.22
2	Tends to find fault with others				-.58	.34
42	Likes to cooperate; goes along with others		.19		.56	-.10

Item		<i>O</i>	<i>C</i>	<i>E</i>	<i>A</i>	<i>N</i>
22	Is generally trusting	.21			.53	-.17
27	Can be cold and distant with others			-.35	-.49	.32
<i>Neuroticism</i>						
19	Worries a lot		-.11	-.20		.76
9	Is relaxed, handles stress well	.13	.24	.14	.17	-.73
14	Can be tense			-.14	-.19	.73
39	Gets nervous easily			-.34		.71
24	Doesn't get easily upset, emotionally stable	.10	.12	.11	.19	-.67
34	Stays calm in tense situations	.11	.38		.11	-.62
4	Is depressed, blue			-.18	-.29	.59
29	Can be moody				-.38	.49
% Variance Explained		9.90	9.90	9.83	9.90	10.21
α		.81	.83	.85	.84	.87

Note: Items Copyright 1991 by Oliver P. John. Reprinted with permission.

Internal Reliability. Internal reliability considers the consistency of items within a test (e.g., how closely related are *Agreeableness* items?). The internal reliability of the BFI-44-Parent was “good”, with Cronbach’s alphas for all factors above .80 inclusive (*Openness* $\alpha = .81$; *Conscientiousness* $\alpha = .83$; *Extraversion* $\alpha = .85$; *Agreeableness* $\alpha = .84$; *Neuroticism* $\alpha = .87$).

Convergent Validity. Convergent validity considers relationships between our instrument and other measures which are expected to be related. Our analyses show that the BFI-44-Parent correlated with the SDQ (Goodman, 1997) scales to an extent that falls in line with previous findings by Muris et al. (2005; who tested 12-17 year olds on the Big Five Questionnaire (BFQ-C). *Openness*, *Conscientiousness*, *Extraversion* and *Agreeableness* all significantly positively correlated with the *Pro-social Behaviour* subscale of the SDQ at the $p < .001$ level with r ’s between .32 and .67 ($n = 538$). *Neuroticism* was significantly related to all difficulties (*Hyperactivity* $r = .30$, *Conduct* $r = .44$, and *Peer Problems* $r = .44$) and was highly correlated (i.e. above .5, Cohen, 1988) with *Emotional Difficulties* ($r = .74$) and *Total Difficulties* ($r = .68$). Conversely, *Total Difficulties* were negatively correlated with *Agreeableness* ($r = -.56$, $p < .001$) and *Conscientiousness* ($r = -.60$, $p < .001$) in particular. Smaller negative correlations (i.e. below .3 Cohen, 1988) were found between *Total Difficulties* and *Extraversion* ($r = -.25$, $p < .001$), and *Openness* ($r = -.27$, $p < .001$). Again these findings are in line with previous literature (Muris et al., 2005).

Further support for the validity of the BFI-44-Parent comes from correlations with the SCARED anxiety questionnaire (Birmaher et al., 1999, 1997). All subscales of the SCARED positively correlated with *Neuroticism* at $p < .001$ (r 's between .46 to .69, $n = 523$) and *Social Anxiety* in particular was highly negatively correlated with *Extraversion* ($r = -.65$, $p < .001$). There were also smaller negative correlations between all anxieties and *Extraversion*, *Agreeableness* and *Conscientiousness* at the $p < .001$ level (r 's between .16 to .45). These findings are broadly in line with Muris et al. (2009) who did not look *within* the SCARED but found a strong positive correlation between the SCARED overall and *Neuroticism*, as well as smaller negative correlations with *Extraversion* and *Agreeableness*. We additionally found, like Muris et al. (2009), a small negative correlation between *Openness* and the overall SCARED ($r = -.16$, $p < .001$). However there was one small difference in which we find a small negative correlation with *Conscientiousness* and the overall SCARED ($r = -.20$, $p < .001$) where Muris et al. (2009) found no correlation.

Lastly we looked at correlations with the Empathy Quotient (EQ; Auyeung et al., 2009). Here we found expected positive correlations, based on Melchers et al. (2016; who tested adults using the NEO-FFI), between the Big Five traits and the EQ, with *Openness*, *Conscientiousness* and *Extraversion*, ($n = 510$, r 's between .33 and .47, $p < .001$) and particularly *Agreeableness* ($r = .68$, $p < .001$). We also found a moderate negative correlation between *Neuroticism* and *Empathy* ($r = -.45$, $p < .001$). These findings were in line with Melchers et al. (2016), except that *Neuroticism* was more strongly negatively correlated with the empathy quotient here.

Discussion

We aimed to validate the BFI-44-parent questionnaire on results from 550 parents of children aged 6-10 years. We found evidence to support the predicted five components. We found good internal reliability, and convergent validity was in line with previous literature. We discuss the implications of these findings more fully in our General Discussion. It is worth noting that we validated this questionnaire using other parent-rated instruments (e.g., SCARED). So whilst this measure is reliable and valid according to a number of parent-reports about their children, it is unclear whether there would be high agreement between parent-ratings and self-reported child ratings. Previous studies have

found only low-to-moderate agreement (Barbaranelli et al., 2003) so we explore this further in the study below. In Study 2 we validate two self-report measures of personality elicited directly from children themselves.

Study 2: Validating children’s self-report measures; The Definitional BFI-44-C and the Pictorial BFI-10-C

Here we aimed to assess two self-report personality measures for children. The first is a written questionnaire (Definitional BFI-44-C) which we have adapted from an existing instrument targeted at older children (BFI-44-A; John, Donahue, & Kentle, 1991; John et al., 2008; John & Srivastava, 1999). The adult version (BFI-44) of this questionnaire had previously been validated on children aged 10+ years by Soto et al. (2008) whose aim was to look at age differences in several measurement properties of youths’ self-reports, but who also found the expected five factors in 10 years olds (i.e., they conducted analyses for each year from age 10 to age 20; Soto et al., 2008). Here, we have added verbal definitions to make the BFI-44-A questionnaire suitable for a yet younger audience, and we test our adapted “Definitional” questionnaire on a cohort of 8-11 year olds. Secondly, we present our new, short, illustrated questionnaire: the Pictorial BFI-10-C, which we validate on children aged 6-10 years.

Methods

Participants. Children were tested in two sessions (See Figure 2). In Session 1 we tested 3349 students from School Years 2 to 5 (age 6-10 years), recruited from the same UK primary schools described in Study 1. All 3349 children completed the Pictorial BFI-10-C ($n = 3349$; mean age 7.92, SD 1.22). Of these children, 1639 were girls (mean age 7.91, SD 1.23) and 1707 were boys (mean age 7.94, SD 1.21). In Session 2, we tested a subset of these children approximately 6-10 months later, when they were now in School Years 3 to 6 ($n = 1279$, mean age 9.00, SD 1.20; see below for our selection criterion). Of these, 641 were girls (mean age 9.06, SD 1.21) and 634 were boys (mean age 8.94, SD 1.19). In Session 2, these children received the Pictorial BFI-10-C again. Finally, an older subset of these children re-tested in Session 2 additionally completed the Definitional BFI-44-C. These were all the re-tested children who were in Year Four or above at the time of

Session 2 (i.e., school Years 4-6, $n = 862$, mean age = 9.66, $SD = 0.85$). This older subgroup comprised 444 girls (mean age = 9.69, $SD = 0.86$) and 418 boys (mean age = 9.62, $SD = 0.83$).

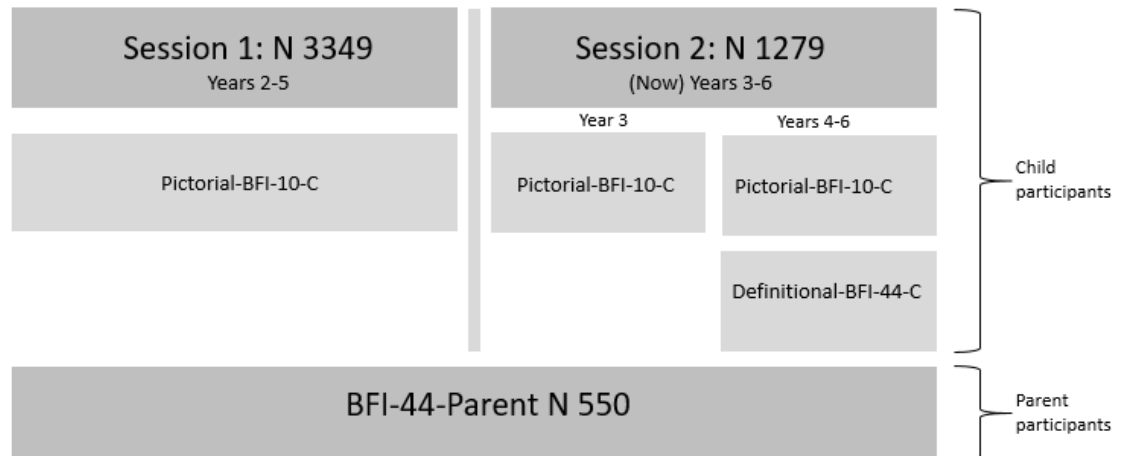


Figure 2: A visual depiction of who was included in each testing session, and which personality tasks they took.

Within the 3349 children tested in Session 1, there were all 550 of the children whose parents were tested in Study 1. Of the children tested in Session 2 who repeated the Pictorial BFI-10-C, there were 228 of the 550 whose parents took part in Study 1 (110 boys, mean age = 8.94 $SD = 1.20$; 118 girls, mean age = 8.98, $SD = 1.14$). Of the children tested in Session 2 who additionally took the Definitional BFI-44-C, there were 152 of the 550 whose parents took part in Study 1 (73 boys, mean age = 9.62 $SD = 0.82$; 79 girls, mean age = 9.61, $SD = 0.82$).

Materials and Procedures. As noted above, we visited schools on two occasions, separated by 6-10 months. On the first occasion, all children were tested on the Pictorial BFI-10-C. On the second occasion, a subset of children were revisited (see below for our selection criteria) and this smaller group were given either just the Pictorial-BFI-10-C (if they were in Year 3 or below), or they were given both the Pictorial-BFI-10-C and the Definitional BFI-44-C (if they were in Year 4 or above). For children given both, half saw the Definitional questionnaire first, and half saw the Pictorial questionnaire first.

Children also completed 12 additional tasks across the two testing sessions depending on their age, and these tests were unrelated to the current study. (The score in one of these tests in Session 1 -- a multisensory learning test -- determined which children were revisited in Session 2 given the aims of our other studies; those revisited were a selection of children who showed below average, average, or above average multisensory learning; see Simner, Alvarez, Rinaldi, et al., 2019; Simner, Rinaldi, et al., 2019).

Children were tested within their classrooms, which had an average size of 25.3 pupils (SD = 5.0, range = 8-32). Each class cohort was tested by three researchers at any given time. After gaining consent from gatekeepers, parents and children, the children were guided through the activities, described below.

The Definitional BFI-44-C. This is an adaptation of the existing questionnaire BFI-44-Adolescent, which presents 44 statements and requires participants to respond on a scale from “Disagree strongly/ Disagree a little/ Neither agree nor disagree/ Agree a little/ Agree strongly”. Each statement relates to one of the Big Five personality traits (ten items for *Openness*, nine for *Agreeableness* and *Conscientiousness*, and eight for *Extraversion* and *Neuroticism*) and begins “I see myself as someone who...” For example, Item 20 relates to the trait of *Openness* and states “I see myself as someone who has an active imagination”. As before, some traits are positively expressed (e.g., *Conscientiousness*; “I see myself as someone who does things carefully and completely”) and some are negatively expressed (“I see myself as someone who can be somewhat careless”).

To adapt the test for our child cohort, we identified 14 late-acquired vocabulary items within the questionnaire which were deemed difficult for our younger sample (8+ years) to comprehend. These words were “fault”, “depressed”, “quarrels”, “reliable”, “tense”, “generates”, “enthusiasm”, “forgiving”, “unorganized”, “imagination”, “trustworthy”, “assertive”, “distant,” and “cooperate”. For each word we added a definition (see below for how definitions were displayed). For instance, for the item, “I see myself as someone who generates a lot of enthusiasm,” we defined the final word as “This means getting excited about things”. Based on the age of acquisition (AoA) database by Kuperman, Stadthagen-Gonzalez, and Brysbaert (2012), the mean AoA of the original 14 items was 8.62 years (SD = 1.35), and the mean AoA of our clarification of these items was 5.87 years (SD = 1.20). Table 2 shows all the terms we defined, and their corresponding AoAs.

Table 2

Age of acquisition (AoA; Kuperman et al., 2012) for added definitions within Definitional BFI-44-C items.

Original Item (AoA for underlined)	Definition (AoA for underlined)
Tends to find <u>fault</u> with others (6.94)	This means often thinking other people do things <u>wrong</u> (4.22)
Is <u>depressed</u> , blue (9.47)	This means being very <u>sad</u> (3.24)
Starts <u>quarrels</u> with others (8.32)	This means someone who starts <u>arguments</u> (7.55)
Is a <u>reliable</u> worker (9.32)	This mean being a <u>hard worker</u> (6.56; hardworking)
Can be <u>tense</u> (9.35)	This means feeling <u>worried</u> (6.65)
Is <u>talkative</u> (8.00)	This means <u>chatty</u> (8.22 ⁴)
Generates a lot of <u>enthusiasm</u> (9.05)	This means getting <u>excited</u> about things (6.21)
Has a <u>forgiving</u> nature (7.28)	This means you <u>forgive</u> people when they do something wrong (5.44)
Tends to be <u>disorganised</u> (10.11; unorganised)	This means being <u>messy</u> (5.05)
Has an active <u>imagination</u> (6.28)	This means liking make-believe and <u>imagining</u> things (6.06; imagine)
Is generally <u>trusting</u>	This means you <u>trust</u> other people (6.55)
Takes charge, has an <u>assertive</u> personality (10.44)	This means you like to be the <u>boss</u> (6.16)
Can be cold and <u>distant</u> with others (8.95)	This means not <u>showing</u> how you <u>feel</u> (6.21; 5.11)
Stays calm in <u>tense</u> situations (9.35)	This means staying calm when things get <u>difficult</u> (5.85)
Likes to <u>cooperate</u> ; goes along with others (8.28)	This means <u>happy</u> to do what other people <u>want</u> (6.17; 4.16)

⁴ Our source study for AoA is based on American English, and was chosen due to its impressive sample size of items and subjects. This source shows one discernible difference from British English: post-hoc norming on 10 speakers of British English shows that the word ‘chatty’ is acquired far earlier by British English speakers (mean = 4.70 years; SD = 1.03).

Although our questionnaires could be administered using pencil-and-paper, we used touchscreen electronic tablets to expedite data coding and analysis. Children were given individual tablets, one per child. These were Acer Aspire SW3-016 tablets or Acer One 10 tablets, running on Intel® Atom TM x5-Z8300 Processors with Windows 10 using 10.1" HD LED IPS (1280 x 800 pixels) Multi Touch Displays. These tablets presented six items per screen over eight screens. Each item appeared adjacent to electronic response-buttons displaying the Likert labels (from left to right: “disagree strongly” to “agree strongly”). Where a definition had been written for vocabulary items, a question mark appeared next to the word which, when clicked, presented the definition as a pop-up (see Figure 3 for a screenshot of the test (top panel) and the pop-up definition (bottom panel)).

I see myself as someone who...

...is helpful and unselfish with others	Disagree strongly	Disagree a little	Neither agree nor disagree	Agree a little	Agree strongly
...can be somewhat careless	Disagree strongly	Disagree a little	Neither agree nor disagree	Agree a little	Agree strongly
...is relaxed, handles stress well.	Disagree strongly	Disagree a little	Neither agree nor disagree	Agree a little	Agree strongly
...is curious about many different things	Disagree strongly	Disagree a little	Neither agree nor disagree	Agree a little	Agree strongly
...is full of energy	Disagree strongly	Disagree a little	Neither agree nor disagree	Agree a little	Agree strongly
...starts quarrels with others ?	Disagree strongly	Disagree a little	Neither agree nor disagree	Agree a little	Agree strongly

I see myself as someone who...

...is helpful and unselfish with others	Disagree strongly	Disagree a little	Neither agree nor disagree	Agree a little	Agree strongly
...can be somewhat careless	Disagree strongly	Disagree a little	Neither agree nor disagree	Agree a little	Agree strongly
...is relaxed, handles stress well	Disagree strongly	Disagree a little	Neither agree nor disagree	Agree a little	Agree strongly
...is curious about many different things	Disagree strongly	Disagree a little	Neither agree nor disagree	Agree a little	Agree strongly
...is full of energy	Disagree strongly	Disagree a little	Neither agree nor disagree	Agree a little	Agree strongly
...starts quarrels with others ?	Disagree strongly	Disagree a little	Neither agree nor disagree	Agree a little	Agree strongly

This means someone who starts arguments

OK

Figure 3. A screenshot of the Definitional BFI-44-C, the top panel shows one page of the questionnaire and the bottom panel shows the pop up that appears when the question mark

is clicked to provide a definition; in this case for the item “I see myself as someone who starts quarrels with others”.

Children were given the following instructions, “You’re going to read some sentences that might describe you, or they might not. For example ‘I see myself as someone who is bossy.’ Is this true about you?” The responses were explained to the children, and children were shown how to click a question mark if they did not understand a word, and to put their hand up if they still did not understand. Students completed the Definitional BFI-44-C in approximately 10 minutes.

The Pictorial BFI-10-C. This is our novel 10-item personality questionnaire for children, containing two items for each of the Big Five personality traits (*Openness*, *Conscientiousness*, *Extraversion*, *Agreeableness*, and *Neuroticism*). These 10 items were based on the BFI-10 (Rammstedt & John, 2007), which we first rewrote to make its statements more understandable to our young audience. In this, we reapplied vocabulary items with simpler synonyms more suitable for children 6+ years (and also took reference from the BFI-44-A, which overlaps in questions). Based on Kuperman et al’s (2012) age of acquisition (AoA) database we ensured all synonyms were learned before or at 6 years old (See Table 3 for original and adapted items, and AoAs.). We also rewrote complex phraseology where needed (e.g., we changed “I am generally trusting” to “I trust people - they often tell the truth”). In all cases we kept the meaning faithful to the original BFI-10. Finally, since our questionnaire would present two images per item, we wrote an alternative version of each statement expressing opposing meaning, with only one word difference where possible (e.g., “I do things carefully and completely” and “I don’t do things carefully and completely”).

Table 3

Original items from the BFI-10 (for adults) and equivalent items from our Pictorial BFI-10-C, with age of acquisition in years (AoA; Kuperman et al., 2012)

Original BFI-10 item (AoA for underlined)	Pictorial BFI-10-C (AoA for underlined)
<u>Reserved</u> ; keeps thoughts and feelings to self (NA)	I don’t tell people what I’m <u>thinking</u> or <u>feeling</u> (4.8; <u>think</u> , 5.3; <u>feeling</u>)
Is generally <u>trusting</u> * (NA)	I <u>trust</u> people – they often tell the <u>truth</u> (6.6; <u>trust</u> , 4.4; <u>truth</u>)
Tends to be <u>lazy</u> (6.4)	I’m often <u>lazy</u> (6.4)

Original BFI-10 item (AoA for underlined)	Pictorial BFI-10-C (AoA for underlined)
Is <u>relaxed</u> , handles stress well (7.6)	I don't get <u>upset</u> if things go wrong (5.3)
Doesn't like <u>artistic</u> things (plays, music) (6.2; art)	I don't like <u>artistic</u> things, like music or art (6.2; art)
Is <u>outgoing sociable</u> (8.1; outgoing; 10.0; sociable)	I prefer <u>playing</u> with other children (4.1; play)
Tends to find <u>fault</u> in others* (6.9)	I always think other children are doing things <u>wrong</u> (4.2)
Does things <u>carefully</u> and <u>completely</u> (5.1; careful, 6.6; complete)	I do things <u>carefully</u> and <u>completely</u> (5.1; careful, 6.6; complete)
Gets <u>nervous</u> easily (6.8)	I get <u>nervous</u> easily (6.8)
Has an active <u>imagination</u> * (6.3)	I'm good at <u>make-believe</u> and <u>imagining</u> things (NA; make-believe, 6.1; imagine)
Is <u>considerate</u> and <u>kind</u> to almost everyone (8.2; considerate, 4.9 kind)	I'm <u>kind</u> and <u>caring</u> to many people (4.9; kind, 5.7; care)
Likes to <u>think</u> and play with <u>ideas</u> * (4.8; think, 6.0; idea)	I like working out puzzles or hard questions

Note: Items Copyright 1991 by Oliver P. John. Reprinted with permission. Asterisks denote items change due to complex phraseology.

Next we paired each two-statement item (e.g., "I do things carefully and completely" and "I don't do things carefully and completely") with a pair of illustrations. These were drawn by a professional illustrator as 4.8x5.8 cm digital line drawings. One picture depicted the positive valence statement (e.g., "I do things carefully and completely") and the other picture depicted the opposing statement ("I don't do things carefully and completely"). Each pair of pictures matched in every detail (e.g., setting, activity) except for the personality trait under investigation. For instance, the item, "I prefer playing with others," and its opposing statement, "I prefer playing on my own," show the same play-activity (i.e., model-building; see Figure 4 for example). This was to ensure children made choices based on personality rather than extraneous factors. Each picture depicted an equal number of boys and girls, aside from the central protagonist. The protagonist was designed to be gender-neutral, and was selected from a shortlist of seven candidate illustrations, normed on n=12 adults who rated the appearance of each illustrated child on

a scale from 1 (very female) to 7 (very male). Mean ratings ranged from 2.8 to 5.3, with the winning illustration being closest to neutral, with the lowest variance and a mean rating of 3.7 ($SD = 1.2$, where neutral is 3.5).


Which one is like you?

Completely ☐

Mostly ☐

Sometimes ☐

Just a bit ☐




I prefer playing with other children

☐ Completely

☐ Mostly

☐ Sometimes

☐ Just a bit



I prefer playing on my own

Figure 4: Example items measuring *Extraversion* on The Pictorial BFI-10-C

The original BFI-10 has two items for each component of the Big Five, one negatively loading and one positively loading onto the component (e.g., *Conscientiousness*: “Does things carefully and completely” and “Tends to be lazy”). Within our pictorial questionnaire, both these items now have opposing valence statements (“I do/ don’t do things carefully and completely” and “I’m often/ not often lazy”). Within our items, we presented on the left whichever statement originated from the BFI-10 (e.g., “I do things carefully and completely”), and presented the other on the right (e.g., “I don’t do things carefully and completely”). Items were then presented in a fixed, pseudo-randomised order to ensure that items from the same component would not be consecutive (see *Appendix A* for full questionnaire and instructions).

As noted above, children were tested on the Pictorial BFI-10-C twice, separated by 6-10 months. As before, our questionnaires could be administered using pencil-and-paper or touchscreen electronic tablets. Tablet testing was used by all children in the second round of testing. Tablet testing was also used for 48 classes in the first round of testing. For the remaining 81 classes in round one, the questionnaire was presented as a pencil-and-paper task (because our on-screen app was still in development). The electronic and paper tests were identical in appearance and design and there were no significant differences in our

data between the two test presentations⁵. In the paper version there were three items per page; in the electronic version, each item was shown screen by screen with an arrow button in the bottom right hand corner to advance to the next screen. The app prevented children from choosing more than one box, while this role was undertaken by the supervising researcher for pencil-and-paper tests. Within the tablet version, if children tried to move onto the next screen without answering, response-options flashed twice; if children pressed one of the pictures instead of a response, responses again flashed twice with a prompt of “how MUCH is it like you?” Again, these procedures were enforced electronically for the tablet version, and by intervention from the researchers for the pencil-and-paper version.

Children were given the following instructions (See Appendix A for full wording): “Please look at the pictures of the children below and read the words saying what the children are like. On one side is a different kind of child to the other side. For each one, choose which side is most like you! Then, once you’ve chosen which side is like you, click a box to say whether it’s only *just a bit* like you, *sometimes* like you, *mostly* like you, or *completely* like you.” Children completed the Pictorial BFI-10-C in approximately 5-10 minutes.

Results.

Ten children were excluded from the Pictorial BFI-10-C analyses. Nine children were flagged by teachers as being newly arrived in the UK with very low levels of English (5 male, 4 female) and one female was removed because she was too old for her year group. These children were not a part of the Definitional BFI-44-C sample, so no children were removed from this latter cohort.

In our analyses we look for validity in our tools but also account for a common response bias found in childhood testing. If a participant consistently agrees or disagrees with items which are logical opposites (e.g., “is talkative” and “tends to be quiet”) this is taken as evidence of an acquiescent response style (Soto & John, 2009; Soto et al., 2008). A

⁵T-tests showed no significant differences between the paper and tablet responses for any of the five factors (*Openness* $p = .84$, *Conscientiousness* $p = .56$, *Extraversion* $p = .56$, *Agreeableness* $p = .07$ and *Neuroticism* $p = .40$)

widely-accepted approach, followed here, is *ipsatization* (Soto & John, 2009; see also Danner, Aichholzer, & Rammstedt, 2015; Rammstedt & Farmer, 2013). In this, a layer of acquiescence is effectively removed from the data to reveal underlying patterns and trends that might otherwise be obscured. Practically speaking, a mean is calculated for each participant across the overall questionnaire, and this mean is subtracted from each item-response, and then the result is divided by the participant's overall standard deviation. This estimates the level of bias for each child and then standardizes scores accordingly. Following Soto and John (2009), we calculated each participant's mean score based on the 32 items in the BFI which have a matched pair; i.e., there are 16 pairs of the type: "is talkative" and "tends to be quiet".

Below we report separate factor analyses based on both the ipsatized and non-adjusted scores, and then compare these two approaches to see whether ipsatized values are superior, suggesting evidence of acquiescence. Again, practically-speaking this is done by comparing component loadings using a Procrustes rotation with an idealized Big Five personality structure (Rammstedt & Farmer, 2013). We took the BFI-44 in US adults reported by Benet-Martínez and John (1998), since this contains the factor analysis output needed for the Procrustes rotation. Where ipsatization shows superior analyses, we suggest ipsatization be used in future applications of our test and we provide syntax files in our *Appendices* (see Appendix B and Appendix C) for future researchers to apply in their own research. In order to compare component structures and in line with our predictions of a five component model, we again follow the approach of Rammstedt and Farmer (2013) and force a five component solution in all of our factor analyses.

Definitional BFI-44-C.

Construct Validity. As in Study 1, we first conducted a principal components analysis on the non-adjusted (i.e. non-ipsatized) 44 items with orthogonal varimax rotation. The KMO was .86 suggesting again we had acceptable sampling adequacy. The five components accounted for 36.03% of the variance. With unadjusted items (i.e. non-ipsatized), all *Openness* items loaded onto one component, and trends suggest two other components representing *Conscientiousness* and *Agreeableness*. However, items belonging to the remaining components were interspersed and there was no clear

emergence of *Neuroticism* or *Extraversion*. See Table 1 in *Supplementary Information* (SI) at the end of the chapter for component loadings.

To determine whether data reflected acquiescence-bias, we next conducted the same principal component analysis with the items ipsatized. We again extracted five components, which accounted for 31.96% of the variance. This time we found evidence suggesting all five components, with only five items failing to converge highest on their expected components (items: 26, 6, 8, 15 and 29). In all five of these cases the expected component was the next highest loading. See Table 4 for the component loadings.

Table 4.

Component loadings of ipsatized Definitional BFI-44-C items. Table shows loadings over .10. Strongest component loadings appear in bold. Item numbers correspond to Definitional BFI-44-C questionnaire. For header, O=Openness; C=Conscientiousness; E=Extraversion; A=Agreeableness; N=Neuroticism.

BFI-44 Item		<i>O</i>	<i>C</i>	<i>E</i>	<i>A</i>	<i>N</i>
<i>Openness</i>						
30	Likes artistic and creative experiences	.67				
25	Is creative and inventive	.66	.11	.17	.12	
44	Knows a lot about art, music, or books	.60	.18			
20	Has an active imagination	.56	-.13			
10	Is curious about many different things	.51		.16		
41	Doesn't like artistic things (plays, music)	-.50	-.14		-.11	-.16
40	Likes to think and play with ideas	.50		.11		
5	Is original, comes up with new ideas	.50	.19	.14		
35	Likes work that is the same every time (routine)	-.22	.17		-.11	
15	Is clever, thinks a lot	.22	.49	.12		-.13
<i>Conscientiousness</i>						
18	Tends to be disorganized		-.59	.14		
28	Keeps working until things are done		.54		.25	-.10
33	Does things efficiently (quickly and correctly)		.53			-.26
13	Is a reliable worker	.11	.53	.12	.30	
43	Is easily distracted; has trouble paying attention		-.44	.13	-.29	
3	Does things carefully and completely	.16	.44	-.17	.15	
23	Tends to be lazy		-.43		-.25	
38	Makes plans and sticks to them		.33	.12		
8	Can be somewhat careless		-.25	.20	-.33	
<i>Extraversion</i>						
6	Reserved; keeps thoughts and feelings to self			-.19		.24
21	Tends to be quiet			-.71	.15	

BFI-44		<i>O</i>	<i>C</i>	<i>E</i>	<i>A</i>	<i>N</i>
Item						
31	Is sometimes shy, inhibited		-.12	-.65		.27
36	Is outgoing, sociable			.47	.20	-.12
1	Is talkative	.20		.45	-.15	
16	Generates a lot of enthusiasm	.17		.43	.30	
11	Is full of energy	.16		.42	.10	
26	Takes charge, has an assertive personality		.12	.30	-.42	
<i>Agreeableness</i>						
37	Is sometimes rude to others		-.24		-.59	
32	Is considerate and kind to almost everyone		.22		.56	
17	Has a forgiving nature				.56	
7	Is helpful and unselfish with others		.11		.49	
12	Starts quarrels with others	-.21	-.11	.21	-.42	.15
42	Likes to cooperate; goes along with others		.10		.37	-.18
2	Tends to find fault with others	-.19		.17	-.37	.11
22	Is generally trusting	.17	.20	.17	.36	
27	Can be cold and distant with others	-.13		-.18	-.36	.13
<i>Neuroticism</i>						
19	Worries a lot			-.16		.71
14	Can be tense					.62
34	Stays calm in tense situations	-.12	.26		.17	-.54
24	Doesn't get easily upset, emotionally stable					-.53
4	Is depressed, blue	-.21			-.12	.49
9	Is relaxed, handles stress well	-.12		.11	.24	-.49
39	Gets nervous easily	.10	-.11	-.52	.14	.46
29	Can be moody		-.23		-.39	.33
% Variance explained		6.90	6.15	5.83	7.01	6.07
α		.68	.70	.66	.73	.68

Note: Items Copyright 1991 by Oliver P. John. Reprinted with permission

To compare ipsatized and non-adjusted component structures, we compared each to an idealized component structure in turn, and then evaluated the performance of each. We looked for congruence between the two component structures (i.e. between the ipsatized component structure and the idealized component structure, and then between the non-adjusted component structure and the idealized component structure) of above .85 according to Lorenzo-Seva & ten Berge's (2006) criteria. They suggest .85 to .94 is 'fair' similarity and that .95 or above is 'excellent' (since this shows that components can be considered equivalent between the current and idealised component structures). We can then conclude whether adjusted or non-adjusted scores represent a better fit. After running a Procrustes rotation on each, we found the ipsatized structure held "fair" agreement with

the ideal component structure (*Openness* = .88, *Conscientiousness* = .86, *Extraversion* = .87, *Agreeableness* = .89, *Neuroticism* = .90) while the non-adjusted component structure was worse, performing below “fair” in all except from one case, suggesting the ipsatized components are more appropriate (*Openness* = .84, *Conscientiousness* = .78, *Extraversion* = .83, *Agreeableness* = .86, *Neuroticism* = .84). In summary, these analyses suggest our results show an acquiescence bias and therefore, the ipsatized scores are more appropriate for future use and are used henceforth in the analyses below.

Internal Reliability. The internal reliability was relatively low, but in line with previous research on children aged 5-7 years assessed using the Berkeley Puppet Interview (Measelle et al., 2005). Cronbach’s alpha showed moderate reliability for all components (*Openness* α = .68, *Conscientiousness* α = .70, *Extraversion* α = .66, *Agreeableness* α = .73, and *Neuroticism* α = .68).

Concurrent Validity. To assess the validity of this instrument we first looked at agreement between parent-rated and child-rated personality. Child-related personality came from the 152 children aged 8-11 who self-reported their personality using this Definitional BFI-44-C, and whose parents had also completed the BFI-44-Parent in Study 1. We found small-to-moderate agreement with all components except for *Agreeableness*, which is broadly in line with previous studies looking at parent-child agreement (Markey et al., 2002). Agreement between child and parent ratings was strongest for *Extraversion* subscales at $r = .45$ ($p < .001$), then *Neuroticism* at $r = .39$ ($p < .001$), then *Openness* at $r = .28$ ($p < .001$), *Conscientiousness* was $r = .31$ ($p = .001$) and *Agreeableness* was non-significant $r = .10$ (*n.s.*).

Convergent Validity. We next looked at convergent validity by comparing children’s self-reported personality subscales against their SDQ (Goodman, 1997), for $n = 149$ children ($n = 75$ girls, $n = 74$ boys) whose parents also completed the SDQ in Study 1 on their behalf. As expected, *Emotional Difficulties* ($r = .36$, $p < .001$), *Conduct Problems* ($r = .21$, $p = .009$) and *Peer Problems* ($r = .32$, $p < .001$) were all positively correlated with *Neuroticism*. *Hyperactivity* was negatively correlated with *Agreeableness* ($r = -.29$, $p < .001$), and *Conscientiousness* ($r = -.27$, $p < .001$), while *Emotional Difficulties* were negatively related to *Extraversion* ($r = -.24$, $p = .003$). These findings are in line with Muris et al. (2005).

We next looked for convergent validity against the SCARED questionnaire (Birmaher et al., 1999, 1997). One hundred and forty-five children's parents had completed the SCARED for their child in Study 1. *Neuroticism* was positively correlated with all of the SCARED's anxiety aspects (r 's between .20 and .41), which is in line with Muris et al.'s (2009) strong correlation between *Total Anxiety* and *Neuroticism*. *Extraversion* was also negatively correlated with all subscales (r 's between -.17 and -.29) and particularly *Separation Anxiety* ($r = -.41, p < .001, n = 145$). This is again in line with Muris et al. (2009); however, they found some additional small correlations between the remaining traits and *Total Anxiety*, which we do not see here.

We lastly looked for convergent validity based on 140 children whose parents in Study 1 had also completed the EQ on their behalf (Auyeung et al., 2009). As expected, *Empathy* was significantly positively correlated with *Agreeableness* ($r = .20, p = .018$) and negatively correlated with *Neuroticism* ($r = -.20, p = .019$). There were no significant correlations between *Openness*, *Conscientiousness*, or *Extraversion* ($r = .05$ n.s., $r = .12$ n.s., $r = -.11$ n.s.). These findings are broadly in line with Melchers et al. (2016), however these latter also found significant positive correlations with *Openness*, *Conscientiousness* and *Extraversion* and the EQ.

The Pictorial BFI-10-C.

We now turn to our second self-report questionnaire, the Pictorial BFI-10-C analyses. We use data from Session 1 unless otherwise stated.

Construct Validity. We first looked at all year groups (Years 2-5) together and again extracted five components from a principal components analysis, with varimax rotation. We found a KMO of .72, which is described as “middling” sampling adequacy by Hutcheson and Sofroniou (1999), which is sufficient to suggest factor analysis is suitable. These five components accounted for 61.51% of the variance. As might be expected with unadjusted scores if there is an acquiescence bias, we could not see the expected Big Five components in the component structure since items did not load together in the expected manner (e.g., all items thought to be associated with *Extraversion* do not load together on a single component). However, there is some convergence suggesting an *Openness* and an *Agreeableness* component. See Table 2 in *SI* at the end of the chapter for component loadings of non-adjusted items.

We therefore next adjusted for a possible acquiescence bias⁶ by looking at the within-person mean centred items and running the same analysis of PCA extracting five components from Varimax rotation. These five components accounted for 66.08% of the variance. Here we now found all five components and all items converged with the highest loading on the expected component except for one *Extraversion* item which converged with *Conscientiousness* (Item six; see *Appendix A* for full questionnaire). See Table 5 for component loadings.

Table 5.

Component loadings of ipsatized Pictorial BFI-10-C items across all children (6-10 years, mean = 8.42). Table shows loadings over .10. Strongest component loadings appear in bold. For header, O=Openness; C=Conscientiousness; E=Extraversion; A=Agreeableness; N=Neuroticism.

Pictorial BFI-10-C Item	<i>O</i>	<i>C</i>	<i>E</i>	<i>A</i>	<i>N</i>
<i>Openness</i>					
Imagine	-.95				
Artistic	.37	.24	.11	.36	.30
<i>Conscientiousness</i>					
Careful		-.87	.16		
Lazy		.57	.30	.14	
<i>Extraversion</i>					
Play			-.96	-.11	
Feeling	.32	.44	.28		-.23
<i>Agreeableness</i>					
Doing Wrong	.12			.86	
Trust		-.34	-.10	-.65	.11
<i>Neuroticism</i>					
Upset		.16	.14	-.18	.77
Nervous		.20	.20		-.76
% Variance explained	11.82	15.28	11.99	13.63	13.35

As before we looked for evidence of acquiescent responding. There was some evidence of acquiescence bias in this data set, perhaps especially in the *Conscientiousness* component which showed improvements with the adjusted data (Non-adjusted: *Openness*

⁶ We used a slightly different method in place of ipsatization given that the Pictorial BFI-10 does not have directly matched items (of the type “is talkative” and “tends to be quiet”). We therefore centered items within participants by subtracting the person mean from *every* item (rather than matched items; Soto et al., 2008).

= .88; *Conscientiousness*= .66; *Extraversion* = .85; *Agreeableness* = .94; *Neuroticism* = .92; adjusted: *Openness* = .89; *Conscientiousness*= .88; *Extraversion* = .93; *Agreeableness* = .95; *Neuroticism* = .93). Therefore, the adjusted data show better similarity to the ideal BFI component structure, with all meeting the .85 similarity threshold.

In order to ensure that the Pictorial-BFI-10-C is appropriate across all children aged 6-10 years, we consider separately younger children (aged 6-7) and older children (aged 8+). In order to split our group into younger and older responders, we chose the youngest age that the Definitional BFI-44-C had been validated on in the previous analysis (i.e. 8 year olds) resulting in an older group of 8-10 year olds, which we compared to a younger group of 6-7 year olds. Given that we found a marginal improvement using the ipsatized items, we conducted a principal components analysis (PCA) using ipsatized items, again outputting five components with varimax rotation for younger children (aged 6-7) and older children (aged 8+) separately. As we found above for the Definitional BFI-44-C, the typical Big Five structure was found with the older group. All items loaded onto the expected component (See Table 6 for component loadings) and “excellent” agreement was found with the ideal component structure in all cases except *Openness*, which just fell short of the .85 “fair” threshold for similarity (*Openness* = .84; *Conscientiousness*= .99; *Extraversion* = .97; *Agreeableness* = .96; *Neuroticism* = .94).

The component structure in the younger group was less robust (see Table 7 for component loadings). This was reflected in the Procrustes rotation with *Agreeableness* and *Neuroticism* showing “fair” agreement with the ideal component structure, but the remaining components being below the acceptable range. This suggests that *Openness*, *Conscientiousness*, and *Extraversion* components in young children were not equivalent to the ideal adult component solution (*Openness* = .72; *Conscientiousness*= .69; *Extraversion* = .72; *Agreeableness* = .94; *Neuroticism* = .95). However, given that we might not expect 6-7 year old children to respond in the same way as adults, we conducted an additional analysis using the Definitional BFI-44-C from children 8-11 years as the ideal solution in the Procrustes analysis. Again we found *Agreeableness* and *Neuroticism* scores were within the “fair” range and *Openness*, *Conscientiousness* and *Extraversion* fell below the .85 threshold (*Openness* = .77; *Conscientiousness* = .81; *Extraversion* = .59; *Agreeableness* = .88; *Neuroticism* = .90). This suggests that children aged 6-7 years do not answer in the same way as older children to questions about their own

Conscientiousness, Extraversion, or Openness. These components should therefore be treated cautiously in young children.

Table 6.

Component loadings from older children aged 8 to 10 (mean = 9.21) with ipsatized items of the Pictorial BFI-10-C Table shows loadings over .10. Strongest component loadings appear in bold. For header, O=Openness; C=Conscientiousness; E=Extraversion; A=Agreeableness; N=Neuroticism.

Pictorial BFI-10-C Item	<i>O</i>	<i>C</i>	<i>E</i>	<i>A</i>	<i>N</i>
<i>Openness</i>					
Imagine	-.90				
Artistic	.43	.23	.13	.30	.34
<i>Conscientiousness</i>					
Careful	-.17	-.83			-.11
Lazy		.76	.16	.13	
<i>Extraversion</i>					
Play	.18		-.87	-.15	
Feeling	.39	.20	.58		-.15
<i>Agreeableness</i>					
Doing Wrong				.88	
Trust		-.15	-.27	-.64	
<i>Neuroticism</i>					
Upset	.11		.13	-.18	.79
Nervous	.13		.28		-.75
% Variance explained	12.47	13.87	13.09	13.47	13.58

Table 7.

Component loadings from younger children aged 6-7 (mean = 7.2) with ipsatized items of the Pictorial BFI-10-C. Table shows loadings over .10. Strongest component loadings appear in bold. For header, O=Openness; C=Conscientiousness; E=Extraversion; A=Agreeableness; N=Neuroticism.

Pictorial BFI-10-C Item	<i>O</i>	<i>C</i>	<i>E</i>	<i>A</i>	<i>N</i>
<i>Openness</i>					
Artistic	.12	.13	.30	.57	.13
Imagine	-.92			-.16	
<i>Conscientiousness</i>					
Careful		-.72	-.10		
Lazy	.21	.40	.49		
<i>Extraversion</i>					
Play			-.93	-.15	
Feeling	.48	.53			

Pictorial BFI-10-C Item	<i>O</i>	<i>C</i>	<i>E</i>	<i>A</i>	<i>N</i>
<i>Agreeableness</i>					
Doing Wrong		.16		.83	-.13
Trust	.15	-.70		-.36	
<i>Neuroticism</i>					
Upset		.18	.19	-.17	.84
Nervous		.30	.28	-.18	-.72
% Variance explained	11.80	16.11	13.35	12.73	12.66

Internal Reliability. Cronbach's alpha penalizes for small numbers of items, so is not recommended for two-item components (Eisinga, Grotenhuis, & Pelzer, 2013). We looked therefore at correlations between items within each personality component. We found significant correlations within all components, with these being typically higher than correlations *across* components (See Table 8). Inter-item reliabilities shown here were higher than the average inter-item correlations of the Definitional-BFI-44-C components (which ranged from .18-.22) but lower than the average inter-item correlations of the BFI-44-Parent (which ranged from .31-.45).

Table 8.

A correlation matrix showing the Pictorial BFI-10-C items correlating with each other. Correlations between items on the same component are highlighted in bold.

Item	Doing Wrong	Artistic	Upset	Careful	Feeling	Trust	Imagine	Nervous	Lazy	Play
Doing Wrong										
Artistic	.13**									
Upset	-.12**	.04*								
Careful	-.19**	-.29**	-.14**							
Feeling	-.09**	.08**	-.08**	-.20**						
Trust	-.32**	-.20**	-.06**	.12**	-.27**					
Imagine	-.16**	-.21**	-.08**	.05**	-.21**	.02				
Nervous	-.01	-.07**	-.28**	-.12**	.08**	-.19	-.10**			
Lazy	.11**	.12**	-.03	-.32**	.13**	-.27**	-.18**	.03		
Play	-.18**	-.18**	-.07**	.02	-.25**	.09**	-.06**	-.20**	-.28**	

Next we looked at test-retest reliability, for children who completed the Pictorial-BFI-10-C in both Session 1 and Session 2 ($n = 1275$). As expected we found moderate correlations all significant at $p < .001$ (*Openness* $r = .43$; *Conscientiousness* $r = .49$; *Extraversion* r

$=.37$; *Agreeableness* $r = .31$ and *Neuroticism* $r = .46$). Given the relatively long retest interval (6-10 months), our results are in line with previous findings (see Measelle et al., 2005 who looked at children 5-7 years using Bekeley Puppet Interview).

Concurrent Validity. We assessed concurrent validity – how well our instrument converges with other personality measures by looking at correlations between the Pictorial BFI-10-C and both the parent and child versions of the BFI-44 (BFI-44-parent; Definitional BFI-44-C). We took the Definitional BFI-44-C and the Session 2 Pictorial-BFI-10-C data (i.e., measured at the same time point) and found small-moderate agreement between the two. *Openness* correlated at $.47$ ($p < .001$), *Conscientiousness* $.43$ ($p < .001$), *Extraversion* at $.20$ ($p < .001$), *Agreeableness* at $.33$ ($p < .001$), and *Neuroticism* $.37$ ($p < .001$) as expected.

We next looked at agreement between the Pictorial BFI-10-C and the BFI-44-Parent and found low but significant agreement for *Conscientiousness* ($r = .20$, $p < .001$), *Neuroticism* ($r = .16$, $p < .001$) and *Openness* ($r = .16$; $p < .001$) but no significant relationship for *Extraversion* ($r = .05$; *n. s.*) or *Agreeableness* ($r = .07$; *n. s.*). This is perhaps unsurprising given that the Pictorial BFI-10-C is not an identical questionnaire, and low agreement between parent and child-ratings of personality is well documented in the literature (Markey et al., 2002). We further investigated this using our older and younger groupings and found that older children ($n = 460$) were marginally more in line with parents (*Openness* $r = .23$, $p < .001$, *Conscientiousness* $r = .24$, $p < .001$, *Extraversion* $r = .05$; *n. s.*, *Agreeableness* $r = .10$, $p = .035$, *Neuroticism* $r = .19$, $p < .001$) whereas 6-7 year olds ($n = 213$) were less in line with parent estimates (*Openness* $r = -.02$, *n. s.*, *Conscientiousness* $r = -.15$; $p = .027$, *Extraversion* $r = .01$; *n. s.*, *Agreeableness* $r = .01$; *n. s.*, *Neuroticism* $r = .10$; *n. s.*) although the sample here was smaller.

Convergent Validity. As before, we sought evidence of convergent validity for our questionnaire (whether our instrument correlates as expected to other related measures) by comparing personality components from the Pictorial BFI-10-C against those from the SDQ, SCARED and EQ (Auyeung et al., 2009; Birmaher et al., 1999, 1997; Goodman, 1997).

For $n = 522$ children whose parents completed the SDQ in Study 1, we found significant positive correlations between *Prosocial Behaviour* and the all components of the Pictorial

BFI-10-C: *Openness* ($r = .18, p < .001$), *Conscientiousness* ($r = .18, p < .001$), *Extraversion* ($r = .23, p < .001$), *Agreeableness* ($r = .10, p = .028$) and *Neuroticism* ($r = .11, p = .011$). *Emotional Problems* were significantly correlated with *Neuroticism* ($r = .14, p = .002$) and there were negative correlations between, on the one hand, *Agreeableness*, and on the other, both *Peer Problems* and *Hyperactivity* in particular ($r = -.10, p = .018, r = -.12, p = .008$ respectively). *Hyperactivity* was also negatively correlated with *Conscientiousness*, *Extraversion*, and *Openness* ($r = -.15, p = .001, r = -.13, p = .002, r = -.10, p = .021$). These findings are broadly in line with Muris et al. (2005).

For $n = 507$ children whose parents also completed the SCARED in Study 1, we found *Neuroticism* positively correlated with *Total Anxiety* ($r = .14, p = .002$), *Social Anxiety* ($r = .14, p = .002$), *General Anxiety* ($r = .13, p = .005$) and *School Avoidance* ($r = .12, p = .006$). This is expected based on Muris et al. (2009), who found a strong correlation between the SCARED and *Neuroticism*, and our data supplements this finding by also looking at individual subscales.

For $n = 497$ children whose parents also completed the EQ in Study 1 (Auyeung et al., 2009), we found small positive correlations between *Empathy* and *Openness* ($r = .15, p = .001$), *Conscientiousness* ($r = .20, p < .001$), and *Extraversion* ($r = .18, p < .001$). However, we found no significant correlation between *Empathy* and *Neuroticism* as we had found within the Definitional BFI-44-C ($r = .09, p = \text{n.s.}$), and only a small correlation between *Empathy* and *Agreeableness* ($r = .09, p = .038$). We did however find the expected correlations with *Openness*, *Extraversion* and *Conscientiousness* based on Melchers et al. (2016) which had been missing with the Definitional BFI-44-C. Conversely, we found no correlation between *Empathy* and *Agreeableness*, or *Empathy* and *Neuroticism*, although these had been found previously by both Melchers et al. (2016) and in our Definitional BFI-44-C above.

Discussion

In this study we presented two self-report questionnaires for children in middle childhood. Our Definitional BFI-44-C provided definitions to augment a previous questionnaire (BFI-44-A) in order to make it suitable for younger children, and we validated this on 8-11 year olds. Our measure showed the expected five component structure after controlling for acquiescence, and had acceptable internal reliability, in line with previous literature

on children (Measelle et al., 2005). We found low agreement with parents when we compared the BFI-44-Parent results to the Definitional BFI-44-C, again in line with previous research (Markey et al., 2002). We additionally found expected convergent validity between our child self-report measure and parent-completed questionnaires: with Goodman's Strengths and Difficulties (SDQ; Goodman, 1997), the Empathy Quotient (Auyeung et al., 2009) and the SCARED (Birmaher et al., 1999, 1997). Again this mirrors previous research (Melchers et al., 2016; Muris et al., 2009, 2005).

We also presented a second self-report questionnaire, the Pictorial BFI-10-C. This is our novel ten-item questionnaire aimed for children aged 6-10 years, which illustrates personality traits in pictorial form. We found this measure to show the expected five component structure when we controlled for acquiescence across the entire age group of 6-10 years. We also found this component structure for children aged 8-11. For younger children only (6-7 years), the five-component structure did not emerge, and although *Agreeableness* and *Neuroticism* were consistent with the ideal solution (from adults or older children), this was not true for the remaining components. Nonetheless, we found the Pictorial BFI-10-C held good levels of test-retest reliability, as well as good levels of concurrent validity with the Definitional BFI-44-C. Additionally, we found expected convergent validity with other measures, including the Goodman's Strengths and Difficulties questionnaire (SDQ; Goodman, 1997) and the SCARED (Birmaher et al., 1999, 1997), again mirroring previous research (Melchers et al., 2016; Muris et al., 2009, 2005). Although we acknowledge that convergent validity with other measures was weaker for the Pictorial-BFI-10 than for the Definitional-BFI-44-C, this is likely due to the shorter length of test.

General Discussion

In this paper we have validated the existing BFI-44-Parent along with two novel or adapted self-report questionnaires: the Pictorial BFI-10-C and the Definitional BFI-44-C. Our data came from very large samples; in some cases, from several thousand children. In the past it has been rare to collect self-report data from children under 12 years, and certainly very rare to collect younger than 10 years (Soto et al., 2008). The first key conclusion of this study, therefore, is that relatively young children in middle childhood

can complete a self-report personality questionnaire, if the language is age-appropriate and/or pictures are provided to make abstract ideas more concrete.

Our child participants aged 8-11 years successfully completed both the Definitional BFI-44-C and the Pictorial BFI-10-C, with the latter being much faster to administer. Received wisdom is that longer questionnaires will always give more robust assessment of personality (e.g., Rammstedt & John, 2007), and we certainly find areas to support this conclusion in our own data. However, our shorter questionnaire has surprising strengths, which may have arisen from its pictorial design. We compared our Definitional (long, written) and Pictorial (shorter) questionnaires in four ways, and the longer questionnaire was arguably superior in internal reliability and concurrent validity. However, when comparing both questionnaires to the idealized component structure after addressing acquiescence, we found that the components of the shorter pictorial questionnaire (with congruence between .84 and .99) were equivalent or even marginally superior to those of the longer one (congruence .84 to .90). Furthermore, the shorter questionnaire allowed for known convergent validity to emerge with the EQ (in *Openness*, *Conscientiousness*, *Extraversion*) that did not emerge in the longer Definitional questionnaire. Nonetheless, the longer questionnaire showed stronger correlations with other convergent measures (SDQ, SCARED) and had stronger internal reliability and concurrent validity (noted above). With ample time we therefore recommend the Definitional BFI-44-C with children aged 8-11, but the shorter Pictorial test remains a viable option when time is limited or when reading demands need to be minimized.

Our finding that children aged 8-11 years show the Big Five components when self-reporting personality (in Definitional and Pictorial questionnaires) are in line with Maćkiewicz and Cieciuch (2016), who validated a Polish-language pictorial personality questionnaire for Polish children with an average age of 9.25 years. With our very youngest children however (aged 6-7 years), we found reliable self-reporting only for *Agreeableness* and *Neuroticism*, notwithstanding our questionnaire's child-friendly pictorial design. The other three components showed a high degree of cross-loadings and this is in line with previous results by Soto et al. (2008). This poorer performance in very young children may relate to greater acquiescence (which we attempted to control for) and/or poor discrimination between components in a conceptual sense. These young children may still think of people as being either "good" or "bad" rather than showing the nuance between components that emerges during later childhood (Soto et al., 2008). This

fact may help to explain why the two components that were robustly elicited in our youngest children (6-7 years) were *Agreeableness* and *Neuroticism*, which are arguably easiest to separate into binary categories of “good” and “bad” compared to other components. The other three components (*Openness*, *Conscientiousness*, *Extraversion*) did not emerge in the self-reports of 6 and 7 year olds, which fits with known difficulties eliciting some of these domains (e.g., *Openness*; Caspi et al., 2005). In summary, although we successfully validated our Pictorial-BFI-10-C in children age 8-11 years, we advise caution when using it to gather self-report data from 6-7 year olds -- especially when interpreting their components of *Conscientiousness*, *Openness* and *Extraversion*. Best practice may therefore be to gather parent data for children aged 6-7 years, in combination with self-reports from the Pictorial-BFI-10-C for *Agreeableness* and *Neuroticism* only.

A further finding from our study relates to acquiescent response styles. We found it crucial to control for this behaviour for our child responders in both questionnaires. This finding is in line with research by Soto et al. (2008) who found that the Big Five structure did not emerge from the BFI-44 in a sample of ten-year-olds without first controlling for acquiescence. Our findings add further weight to the suggestion that acquiescence becomes more important to control as the sample gets younger (or perhaps for samples that are less educated; Rammstedt & Farmer, 2013).

The next key finding from our study relates to our validation of the BFI-44-Parent which, to our knowledge, had not been purposefully validated previously. Our data here provide overall support for this instrument. We found that it has construct validity for the Big Five components, with “good” internal reliability. It also showed expected convergent validity with three other parent-rated questionnaires (SDQ, SCARED, EQ). These findings mirror the validations found previously for the adult self-report version of this same test (BFI-44; John & Srivastava, 1999), but it has been important here to extend this validation to a parent population, giving reports about their children. Our final finding was that, like others (e.g., Markey et al., 2002), we show relatively low agreement between parent-rated personality and child-rated personality. These findings suggest that there are aspects of children’s personalities that are captured by the self which cannot be captured by another person. Developmental researchers might therefore aim to capture both self-rated and parent-rated personality where possible. Tellingly, although parent-rated personalities offered strong convergent validity with other *parent* measures (SDQ, SCARED, EQ), children rating their own personalities did so consistently across different *child*-rated

instruments (Definitional BFI-44-C, and Pictorial-BFI-10-C). Indeed, the correlation across the two child self-report questionnaires in this study were as high as $r=.47$ for some components. Given these differences, a more accurate assessment of personality may be to consider both child and parent viewpoints.

We note a final limitation of our study from testing children in a group setting of approximately 25 students at a time. This may have contributed to the Pictorial BFI-10-C's relatively low levels of retest reliability over 6-10 months (although this is also to be expected for a two-item measure; Eisinga et al., 2013). Nonetheless, it is likely that these reliability estimates would improve if the measure was given in a one-to-one setting.

In conclusion, we have presented three Big Five inventories that might allow developmental researchers to assess the personalities of relatively young children. In deciding which test to use, we recommend that researchers weigh up the strengths and weaknesses of each test while considering testing environment, age of participant, and the particular areas of personality under investigation.

Chapter 2: Supplementary Information

Non-adjusted factor loadings for the Definitional-BFI-44-C

Table 1

Factor loadings of non-adjusted Definitional-BFI-44-C items

BFI-44 Item		1	2	3	4	5
30	Likes artistic and creative experiences	.69			.14	
44	Knows a lot about art, music, or books	.67			.15	
25	Is creative and inventive	.66			.14	.24
5	Is original, comes up with new ideas	.56		-.10	.22	.11
20	Has an active imagination	.54				.24
10	Is curious about many different things	.53				.17
40	Likes to think and play with ideas	.52			.12	.28
41	Doesn't like artistic things (plays, music)	-.49	.28			
1	Is talkative	.33	.30	-.20	-.12	.14
37	Is sometimes rude to others		.64			-.26
12	Starts quarrels with others	-.12	.57			
8	Can be somewhat careless		.55			
43	Is easily distracted; has trouble paying attention		.54		-.26	
26	Takes charge, has an assertive personality	.12	.52		.23	
29	Can be moody	.11	.50	.26	-.19	-.13
2	Tends to find fault with others	-.12	.46	.10		
27	Can be cold and distant with others		.39	.33	.15	
23	Tends to be lazy		.39		-.25	
39	Gets nervous easily	.11		.72		
31	Is sometimes shy, inhibited			.71		
19	Worries a lot	.11	.13	.68	-.13	
21	Tends to be quiet	-.14	-.21	.60	.23	-.12
14	Can be tense		.24	.55		.11
4	Is depressed, blue	-.12	.26	.46	-.13	
6	Reserved; keeps thoughts and feelings to self		.12	.42		
33	Does things efficiently (quickly and correctly)	.18			.63	
28	Keeps working until things are done	.17	-.16		.56	.30
34	Stays calm in tense situations		-.11	-.26	.53	.20
15	Is clever, thinks a lot	.33			.47	.13
13	Is a reliable worker	.26	-.21		.45	.33
3	Does things carefully and completely	.25	-.25	.16	.44	
18	Tends to be disorganized		.30		-.37	.26
38	Makes plans and sticks to them	.14			.34	.14
24	Doesn't get easily upset, emotionally stable		.12	-.30	.30	

BFI-44 Item		1	2	3	4	5
35	Likes work that is the same every time (routine)	-.11	.17	.18	.29	
17	Has a forgiving nature		-.22	.16		.60
16	Generates a lot of enthusiasm	.28				.56
32	Is considerate and kind to almost everyone	.18	-.34	.13	.28	.47
42	Likes to cooperate; goes along with others		-.15		.24	.44
7	Is helpful and unselfish with others	.13	-.27		.13	.43
36	Is outgoing, sociable	.19	.15	-.23	.22	.43
11	Is full of energy	.21		-.23		.41
22	Is generally trusting	.24	-.20		.22	.41
9	Is relaxed, handles stress well			-.19	.36	.40
Eigenvalues		6.10	3.40	2.93	2.00	1.43
% Variance		13.85	7.72	6.66	4.54	3.25

Note: Strongest factor loadings appear in bold. Loadings over .10 are reported.

Non-adjusted factor loadings for the Pictorial-BFI-10-C

Table 2

Factor loadings of non-adjusted Pictorial BFI-10-C items

Pictorial BFI- 10-c Item	1	2	3	4	5
Imagine	.83				.21
Artistic	-.59		.18		.18
Careful	.45	-.15	-.15	.37	-.21
Nervous	.15	.71			-.37
Feeling	-.19	.69			
Lazy	-.12	.50	.15	-.34	.26
Doing Wrong			.87	.11	
Trust	.11		-.62	.35	
Play				.88	.12
Upset					.87
Eigenvalues	2.08	1.25	.99	.95	.89
% Variance	20.80	12.46	9.85	9.52	8.88

Note: Strongest factor loadings appear in bold. Loadings over .10 are reported.

Chapter 3

What is the personality profile of a child synaesthete?

Chapter Summary

In Chapter 2, I validated three personality instruments for use with children. Here, I use some of these measures to ask whether child synaesthetes have a particular personality profile, when compared to their peers without synaesthesia. This follows on from adult synaesthete studies that have attempted to ask similar questions about synaesthesia in adults. This chapter is written in paper form and is currently in review as Rinaldi, L.J., Smees, R, Carmichael, D.A., and Simner, J (2019) *What is the personality profile of a child synaesthete?* Manuscript submitted for publication. My experiments focused only on the strongest measures from Chapter 2 (BFI-44-Parent; Definitional-BFI-44-Child), and these were also the locus of significant results when we examined synaesthetes. Focusing on these two measures therefore allowed for a more robust, readable and concise submission for the literature. Note that where additional models were included in a supplementary information in our article submission, here they have been instead provided at the end of the chapter.

Abstract

Previous research into personality and synaesthesia has focused on adult populations and yielded mixed results. One particular challenge has been to distinguish traits associated with synaesthesia, from traits associated with the ways in which synaesthetes were recruited. In the current study we addressed recruitment issues by testing randomly sampled synaesthetes, and we looked particularly at synaesthesia in childhood. Our child synaesthetes were identified by a screening program within 22 primary schools in the South East of England ($n = 3387$; children aged 6 to 11 years old). This identified two types of synaesthete (*grapheme-colour synaesthesia* and *sequence-personality synaesthesia*), and we tested their personalities using both child-report and parent-report measures. We found strong support for synaesthesia being associated with high *Openness to Experience*, a personality trait linked to intelligence and creativity. Both synaesthesia subtypes showed this feature, supporting previous research in adults (Banissy et al., 2013; Chun & Hupé, 2016; Rouw & Scholte, 2016). We additionally found low *Extraversion* in *grapheme-colour synaesthetes* and high *Conscientiousness* in *sequence-personality synaesthetes*. We discuss our results with reference to earlier recruitment issues, and as to how perceptual differences such as synaesthesia might link to trait-differences in personality.

Introduction

Synaesthesia is a rare perceptual or cognitive trait affecting approximately 4.4% of the population (Simner et al., 2006). People with synaesthesia experience unusual colours, tastes, and other sensations when engaged in everyday activities like reading or listening to music (for review, see Simner & Hubbard, 2013). In the current study we focus on two common types of synaesthesia in which reading letters and numbers triggers either colours (*grapheme-colour synaesthesia*; e.g., the synaesthete feels that A is red, 7 is blue) or personifications (*sequence-personality synaesthesia*; e.g., the synaesthete feels that A is outgoing and male; 7 is generous and female; Simner & Holenstein, 2007; Ward, Simner, & Auyeung, 2005). Both forms are widely recognised variants of synaesthesia with known neurological profiles. For example, people with grapheme-colour synaesthesia show altered white matter connectivity in regions associated with colour processing (see Rouw & Scholte, 2007), while people with sequence-personality synaesthesia show differences in regions associated with social processing (see Simner et al., 2016). Sequence-personality synaesthesia is also known as *ordinal linguistic personification* (OLP) synaesthesia and we refer to it using this shorter acronym throughout our paper. In this study, we ask whether children with either form of synaesthesia show differences in their personality profiles. In other words, we ask: what is the personality of a typical child with synaesthesia? This is the first time any study has considered personality differences in children as a result of this unusual trait. We look specifically at differences between randomly sampled child synaesthetes aged 6-10 years, and their matched non-synaesthetic controls.

It may not be surprising if we were to find that synaesthetes have a specific personality profile, since synaesthetes are known to differ from their peers in a number of ways that transcend synaesthesia itself. For example, adult synaesthetes have better memories (e.g., Rothen, Meier, & Ward, 2012), better spatial processing, and increased visual imagery (e.g., Havlik, Carmichael, & Simner, 2015). Additionally, child synaesthetes show faster processing speed (Simner & Bain, 2018) and heightened vocabulary knowledge (Smees, Hughes, Simner, & Carmichael, 2019). Studies have also examined whether there is a particular personality profile associated with synaesthesia, at least in adults. This earlier research focused on the “Big Five” model of personality (Tupes & Christal, 1961), which considers personality as having five component parts, or *factors*. These factors are widely known as *Openness to Experience*, *Conscientiousness*, *Extraversion*, *Agreeableness*, and

Neuroticism (Goldberg, 1990; McCrae & Costa, 1987). The factor of *Conscientiousness* relates to self-discipline and organisation. *Extraversion* is associated with being outgoing and dominant, *Agreeableness* with traits such as empathy and cooperation, and *Neuroticism* describes how much one is anxious versus emotionally stable. Finally, *Openness to Experience* reflects intellectual curiosity, artistic interest, and imagination (Caspi et al., 2005). Previous research has therefore asked whether adult synaesthetes show differences to their peers in their personalities, as captured by the five factor model.

Three seminal studies have looked at personality traits in adults who had similar types of synaesthesia to the ones we examine here (e.g., coloured letters; Banissy et al., 2013; Chun & Hupé, 2016; Rouw & Scholte, 2016). We review these important studies below because we will be conducting similar research on children. Despite a number of differences across these early studies (see below), all converged on one finding at least: that the synaesthetes they tested showed higher *Openness to Experience* compared to their non-synaesthete controls (Banissy et al., 2013; Chun & Hupé, 2016; Rouw & Scholte, 2016). Additional support for this elevated *Openness* in synaesthetes may come, too, from studies of their creativity -- a feature closely tied to *Openness* (John & Srivastava, 1999). Rothen and Meier (2010b) found that grapheme-colour synaesthesia was more prevalent amongst art students compared to controls, and Ward, Thompson-Lake, Ely, and Kaminski (2008) showed that synaesthetes engage more in artistic pursuits (see also Domino, 1989; Rich, Bradshaw, & Mattingley, 2005). Chun & Hupé (2016) also reported that synaesthetes scored higher on *absorption*, a trait related to the enjoyment of imaginative activities. Finally, synaesthetes also scored higher in convergent thinking (Ward et al., 2008, using the Remote Associates Test; Mednick, 1968), a trait linked to creativity and intelligence (both *Openness* features; John & Srivastava, 1999). In sum, studies have shown in multiple ways that synaesthesia may be linked to the trait of *Openness to Experience* – at least for the adult synaesthetes tested in those earlier studies.

However, although the three studies reviewed above converged on elevated *Openness* in synaesthetes, their findings were problematic in several ways. First, their results differed widely on personality factors *other* than *Openness*. So the 81 grapheme-colour synaesthetes tested by Banissy et al. (2013) also showed lower *Agreeableness* than controls. The 89 synaesthetes tested by Rouw and Scholte (2016; with varying forms of synaesthesia including grapheme-colour synaesthesia) scored significantly higher on *Neuroticism*, and they scored low on *Conscientiousness*. In contrast, Chun and Hupé

(2016) found no other Big Five factors aside from *Openness*, when testing their 29 synaesthetes with multiple forms of synaesthesia (again including grapheme-colour synaesthesia). This body of research therefore suggests that whilst there are likely to be personality profiles associated with being an adult synaesthete, it is unclear precisely what those profiles are, which synaesthesias they affect, and whether any trait other than *Openness* could be replicated⁷.

A second question arises over the ways in which these synaesthetes were recruited for study. Banissy et al. (2013) recruited synaesthetes from a cohort who had reached out to the university and agreed to leave their contact details for future synaesthesia studies. But it is reasonable to assume that this type of volunteer might show certain personality traits irrespective of synaesthesia. For example, they may be driven by high levels of intellectual curiosity, a feature that is important for *Openness to Experience*. Importantly, Banissy et al.'s controls were recruited differently (e.g., some were personal acquaintances who took part in response to personal request). Hence self-referred synaesthetes might score higher on *Openness*, simply by virtue of the recruitment method. In Chun and Hupé (2016) and Rouw and Scholte (2016), steps were taken to minimise sampling biases by ensuring participants were recruited similarly. However, neither study included a fully random sample of verified synaesthetes. For example, Chun and Hupé (2016) included descriptions of synaesthesia in their recruitment materials, which might disproportionately attract intellectually-curious synaesthetes (i.e., those high on *Openness* wishing to better understand themselves). And Rouw and Scholte (2016) did not verify synaesthetes with an objective test -- which is an important stage in confirming bona fide synaesthesia (see Simner, 2012; Simner, Mulvenna, et al., 2006). Simner et al. (2006) have shown that a surprisingly high number of self-declared 'synaesthetes' are not synaesthetes at all. This may be due to misunderstanding, or inattention when filling out the questionnaire. Inattention is tied to low Conscientiousness (Caspi et al., 2005; Grieve, 2012), which is one of the traits in those self-referring as synaesthetes for Rouw and Scholte (2016). Another trait found by Rouw and Scholte was high *Neuroticism*, and this

⁷ One final study found no personality differences for synaesthetes whatsoever, but examined a very different type of synaesthete - sequence-space synaesthetes - who view sequences such as days and months as being projected into spatial arrays (e.g., months of the year might be seen in an oval shape). Ward et al. (2018) recruited synaesthetes and controls without obvious bias, but used a very short personality measure with methodological limitations (e.g., see Gosling et al., 2003). Hence, their null effect may stem from their personality measurement, but might also provide the very real suggestion that different synaesthesias bring different personality profiles. We return to this further below.

has been linked with hypochondria and pathologizing (Costa & McCrae, 1987), so again might be higher in a group of ‘synaesthetes’ if they are, at least in part, people without synaesthesia at all.

In summary, establishing the personality traits of rare groups such as synaesthetes poses particular problems if recruitment; (a) informs subjects about synaesthesia during recruitment; (b) relies on self-diagnoses of synaesthesia without an objective test; (c) recruits synaesthetes differently to controls; or (d) accepts sporadic self-referred volunteers from the population at large (as opposed to identifying synaesthetes using large-scale screening methods as we do here; see below). All these methodological choices are widely used in the literature, and are understandable given difficulties in recruiting synaesthetes, but they may have an adverse effect on assessments of personality. In the current study we therefore take a different approach to avoid these issues, by testing personality in synaesthetes identified by wide-scale screening. This screening targeted the student bodies of 22 primary schools ($n = 3387$; children aged 6 to 10 years old). Recruitment captured virtually the entire student body of targeted classes. Parents/children were free to opt-out but very few did (only 1% of our sample), and this allowed us to capture the personalities of synaesthetes and non-synaesthetes while at the same time avoiding the recruitment problems in adult studies described above.

Aside from the methodological issues discussed above, it is also unclear from adult studies whether different personality traits might relate to different forms of synaesthesia. Previous studies suggest the very real possibility that different forms of synaesthesia might generate different personality profiles. We therefore tested here whether personality is different across two different types of synaesthesia, to examine directly whether variants of synaesthesia are associated with different profiles. One final issue arising from adult studies is that they cannot establish whether personality differences emerge slowly over time, or whether they are observable even in very young synaesthetes. In the current study we therefore examine personalities while synaesthetes were still young (ages 8-11 years). By targeting this age group, we can better understand whether personality differences arise from some *a priori* (e.g., neurodevelopmental) source – emerging early – or whether they arise from repeated exposure to synaesthesia over time – emerging only in adults. For example, repeated exposure to synaesthetic colours might drive synaesthetes to want to engage in creative activities (e.g., painting) and thereby heighten their trait of *Openness* (see Simner, 2019, for a similar account in a first-person

anecdotal report). Here we may not expect peaks of *Openness* in young synaesthetes, given their fewer synaesthetic experiences compared to adult synaesthetes.

In testing the personalities of child synaesthetes, there are several key considerations. Personality must be measured carefully, since traits can be unstable in children and the trait of *Openness* is particularly variable in measurements (Caspi et al., 2005). Moreover, reliability between child-report and parent-report is typically moderate only (Markey, Markey, Tinsley, & Ericksen, 2002; see also McCrae & Costa, 1987). This means that children's self-rated reports may hold additional information, or that children may have a different viewpoint compared to their parents. For this reason, it will be beneficial to use children's own self-report in conjunction with adult ratings, to get a comprehensive assessment of their personalities. Rinaldi, Smees, Carmichael and Simner (2019a) found that children as young as 8 years old can self-report personality on a questionnaire, but children *younger* than 8 struggle to do this. We therefore measure personality using parent-report for children 6-10 years, but also child-report measures for children aged 8 years and older (Rinaldi, Smees, Carmichael, et al., 2019a).

In testing for synaesthesia, we use the *gold standard* method to identify a key marker of synaesthesia known as, *consistency-over-time*. When synaesthetes describe their associations (e.g., A is red, 9 is outgoing) and repeat these descriptions later, they do so with high consistency. Hence the colour of any particular letter (e.g., A is red) does not change markedly over time for any given grapheme-colour synaesthete, and the personality does not change (e.g., 9 is outgoing) for an OLP synaesthete. Diagnostics for synaesthesia therefore elicit associations twice and assess consistency: synaesthetes are identified as those who are extremely consistent over time, while non-synaesthetes are *inconsistent*. One particular challenge in testing child synaesthetes, however, is that their consistency grows with age. At age 6-7 years, child grapheme-colour synaesthetes have only approximately 34% of their alphabet with fixed synaesthetic colours (rising to 71% by age 10-11 years; Simner & Bain, 2013). For this reason, we used an in-house test of consistency that takes into account the rising levels of consistency within child synaesthetes as they age, and sets the diagnostic threshold between synaesthetes and non-synaesthetes accordingly (see Methods, and Simner, Alvarez, Rinaldi, Smees, & Carmichael, 2019; Simner, Rinaldi, et al., 2019).

In summary, here we screen a very large sample of children (aged 6-10 years) for two types of synaesthesia (OLP and grapheme-colour synaesthesia), and at the same time, we measure their personality traits. We have four aims. First, we ask whether synaesthetes have higher *Openness* than their peers, when avoiding the recruitment issues of adult studies. Second, we also seek any other differences in personality profile (higher *Neuroticism*, lower *Conscientiousness*, and lower *Agreeableness*) as found in Rouw and Scholte (2016), and Banissy et al. (2013). Third, we compare our child findings to earlier adult studies, to detect possible developmental differences (see Discussion). Finally, we aim to compare childhood grapheme-colour synaesthesia and OLP synaesthesia, to ask whether different personality traits are tied to different forms of synaesthesia.

Methods

Participants

We tested 3387 children from 22 UK primary schools in East and West Sussex, Southern England, who were aged 6 to 10 years during the first of the two sessions required for this study (see below). Our cohort comprised 1650 girls (mean age = 8.43, SD = 1.17) and 1737 boys (mean age = 8.43, SD = 1.17). Our tests below will divide these children into target groups of synaesthetes and matched controls (see Materials and Procedures for how groups were categorised, and see Results for the numbers within each group).

One hundred and thirty additional subjects were excluded, 40 of whom were opted out either by their parents or themselves (only 1.08% of children across the 22 target schools); nine did not speak English (i.e., were newly arrived in the UK); one was out of her year group in age; and 80 had missing data (e.g., were taken out of class during testing, experienced a technical failure). We also invited the parents of the entire child cohort to take part in our parent-questionnaire. Two hundred and seventy-eight parents of our target children participated (i.e., these 278 were parents of children we subsequently categorised as either synaesthetes or their matched controls; see Results for numbers within each group). This study was approved by the Sussex University Science and Technology Ethics Committee.

Materials and Procedures

Diagnostic for Grapheme-Colour Synaesthesia.

Our in-house test for grapheme-colour synaesthesia in children is reported in detail by Simner, Rinaldi, et al. (2019)⁸. The test was delivered via an app, installed on touch screen tablets handed out, one per child (either Acer Aspire SW3-016 tablets or Acer One 10 tablets, running on Windows 10 with an Intel® Atom TM x5-Z8300 Processor and 10.1" HD LED displays; 1280 x 800 pixels). During the test, children saw 36 graphemes (letters A-Z; numbers 0-9) displayed on-screen, one by one. To the right of the grapheme was a colour palette with 25,600 different colours (see Simner, Rinaldi, et al., 2019 for the design-features which ensured this palette was child-friendly). Children were instructed to “choose the best colour” for each letter or number; they were told there was no wrong or right answer but that they should avoid repeatedly choosing the same colour for everything. Across the entire test, graphemes were presented three times each in a block design, which first randomised A-Z and 0-9 in Block 1, and then pseudo-randomised these 36 graphemes again in each of two more blocks to ensure the same grapheme would never be repeated consecutively. See Appendix D for a screenshot of the grapheme-colour interface.

Following Simner, Rinaldi, et al. (2019), our analysis will compare the three colours given for each grapheme (e.g., the three colours given for the letter A), to assess how consistently each child gave colours for letters and numbers. Children with a large number of highly consistent graphemes were identified as *potential synaesthetes* (see *Results* for the level of consistency required) and these potential synaesthetes were re-tested in a second session 6-10 months later (mean = 7.62 months; SD = 1.12). As well as re-testing these potential synaesthetes ($n = 333$), we also re-tested a group of average controls ($n = 663$). Controls had been matched to potential synaesthetes (in a 2:1 ratio) according to age and sex, but were children who had not shown high consistency within Session 1 (specifically, their consistency in Session 1 had fallen below a threshold placed at 1SD above the mean). Controls were matched from the same school if this was possible, or

⁸ Simner, Rinaldi, et al. (2019) is a methodological paper introducing the synaesthesia diagnostics used here. Simner and colleagues present in-depth details of the testing interfaces (e.g., motivations for design-choices) and of scoring protocols (e.g., describing a variety of ways to compute scores for synaesthetes, and the ways these might suggest synaesthesia at different stages in testing). The current study describes as much detail as the reader requires to ascertain that we have adequately identified synaesthetes.

from a school sharing the same socio-economic status (i.e., using each school's percentage *Free School Meals*, as the UK school-wide benefit linked to low household income; see Taylor, 2018). The number of children retested within each age group, sex group, and experimental group (potential synaesthetes, average controls) is shown in the Results.

In this second session (henceforth *Session 2*; i.e., an average of 7.62 months later), potential synaesthetes were given the same test again, to determine whether they again showed consistency. Only children consistent within Session 1, consistent within Session 2, and consistent longitudinally across sessions (i.e., across 7.62 months) would be ultimately recognised as true synaesthetes (see Results).

Diagnostic for OLP Synaesthesia.

This in-house diagnostic is reported in detail in Simner, Alvarez, et al. (2019). It again tests for synaesthesia by identifying consistent associations, but this time the associations are between graphemes and personifications (e.g., A is a friendly female). In this test, children saw the letters of the alphabet presented in a randomized order down the centre of a page, and were required to match each letter to one of six faces (shown as line drawings). Half the faces were female and half were male, and within each sex, they were either friendly, neutral, or unfriendly. Children were required to choose one face for each letter (e.g., A = friendly female). After completing the task for all letters, children saw the letters again in a re-randomised order 40 minutes later, and they gave their associations again. In other words, they provided two personifications per letter within Session 1. See Appendix E for a screenshot of the OLP interface.

As before, children who showed high consistency within Session 1 (i.e., potential synaesthetes) repeated the test again in *Session 2* which took place 6-10 months later, along with a group of matched controls (who had not been consistent; see Results for how consistency was measured). In Session 1, children completed our test as a pencil-and-paper task but used touchscreen electronic tablets in Session 2 (to expedite scoring). The paper and electronic tests were identical in appearance and design, except that where children drew a line between a letter and its face using a pencil in Session 1, they traced a line with their finger on the touch-screen in Session 2 (and the app drew a line in response). The tablet app prevented children from choosing more than one line per letter,

whereas this role was undertaken by the supervising researcher for the pencil-and-paper version. For the tablet version, children were given the same individual 10" tablets described above.

Personality testing: Child self-report

Children in Session 2 completed a self-report questionnaire called the *Definitional BFI-44-C* (Rinaldi, Smees, Carmichael, et al., 2019a). The items in the questionnaire each relate to one of the Big Five factors of personality, and there were ten items for *Openness*, nine for *Conscientiousness*, and eight for *Extraversion*, *Agreeableness* and *Neuroticism*. For example, one item states, "I see myself as someone who does things carefully and completely" (*Conscientiousness* factor). Children were required to respond on a 5-point Likert Scale from "Disagree Strongly" to "Agree Strongly". This questionnaire is based on the BFI-44 (Big Five Inventory, 44 item; John, Donahue, & Kentle, 1991; John et al., 2008; John & Srivastava, 1999) but provides definitions for words to make the test suitable for children (following Rinaldi, Smees, Carmichael, et al., 2019a, e.g., one items states "I see myself as someone who starts quarrels with others," and has a definition for "quarrel", which appears a pop-up on-screen as "This means someone who starts arguments."). We presented this test during Session 2, using the tablets described above. Since this test is only suitable for older children (8 years and above; Rinaldi, Smees, Carmichael, et al., 2019a), our youngest children did not complete it (i.e., anyone 6-7 years in Session 2).

Personality testing: Parent-report

In order to capture personality in the most comprehensive way possible, we additionally looked at how parents rated the personality of their children, using the equivalent BFI-44 test for parents. The *BFI-44-Parent* (John & Srivastava, 1999) was recently validated by Rinaldi, Smees, Carmichael, et al., (2019a) and is identical to the child-version above, but without definitions, and relates to the child (e.g., "I see my child as someone who... "). Parents completed either a pencil-and-paper version sent by post, or they completed an identical version posted on the website Qualtrics, which they accessed via a URL sent to them by email (the decision of post vs. email was dictated by how each school

communicated with its parents). The questionnaire was sent out during Session 1 testing, and reminder emails were sent during Session 2 testing, and then again once our child-testing was complete.

Results

Identifying grapheme-colour synaesthetes

We diagnosed grapheme-colour synaesthesia according to methodologies advocated by Simner, Rinaldi, et al. (2019). In brief, this involved the following steps. After Session 1, we first identified children who had not followed task instructions (children had been instructed not give the same colour for everything). Here we used a DBSCAN clustering method (Ester, Kriegel, Jorg, & Xu, 1996) to remove large clusters of colours for participants who had, for example, chosen red for all graphemes. This method is described more fully in Simner, Rinaldi, et al. (2019), but essentially recognises large clusters of similarly coloured graphemes, and removes them from all consistency calculations. With this method, we identified and subsequently excluded 30 *potential grapheme-colour synaesthetes* who had large clusters for 40% or more of their graphemes (in either Session 1 or 2). Table 1 summarises our final classification of children at each stage (Session 1 and Session 2 and subject-removal).

After excluding these children, we identified 332 *potential grapheme-colour synaesthetes*, as children who had given consistent colours for their graphemes in Session 1. Specifically, these children had a significantly high number of consistent letters and/or numbers, compared to age-matched peers (i.e., 1.96 standard deviations above the mean for his/her age group; following Simner, Rinaldi et al., 2019). We recognised ‘consistent letters and/or numbers’ by examining the three selections the child had given in Session 1 for each grapheme (e.g., his/her three colours for the letter A). We computed the colour distance between them (in CIELAB colour space; International Color Consortium®, 2004), and where this colour distance was particularly small (1 SD smaller than the mean for that same grapheme across all children) we scored the child 1 point. We then repeated this for all the child’s letters and numbers, thereby giving him/her a *Session 1 Letter Score*

(out of 26) and a *Session 1 Number Score* (out of 10)⁹. We then looked across all children to find the overall distribution of *Session 1 Letter Scores* and *Session 1 Number Scores*. Anyone with a particularly high score (1.96 standard deviations above the mean for their age) showed signs of having many consistently-coloured graphemes. These children were classified as *potential synaesthetes* and were retested in Session 2. (The remaining children were not tested for synaesthesia in Session 2, but 663 of them were paired with *potential synaesthetes* for the purposes of our personality testing; this group were named *average-memory controls*; see Table 1).

In Session 2 we again looked at the consistency of *potential synaesthetes*, in order to identify those who were *true synaesthetes*. We knew that *potential synaesthetes* would have included two types of children: true synaesthetes but also non-synaesthetic children who scored highly within Session 1 simply by chance, by employing some type of strategy (e.g., R = red, G = Green), or from having a superior memory span. We name these *high-memory non-synaesthetes*¹⁰, and the goal of Session 2 was to divide the *potential synaesthete* group into *true synaesthetes* versus *high-memory non-synaesthetes*. True synaesthetes would continue to be consistent when we tested them again, and over a longer period, while *high-memory non-synaesthetes* would not. Hence, we calculated consistency again, but now calculating each child's *Letter Score* and *Number Score* within Session 2. Scores were again out of 26 and 10, respectively, and were computed for each child in the same manner described above (i.e., we scored a point for each letter and number whose colour-distance was below the mean for that grapheme by 1SD or more). Following Simner, Rinaldi et al. (2019), we took our means again from Session 1 because this allowed us to use the largest sample available to set our mean baselines. Using these baselines, we flagged any child whose *Session 2 Letter Score*, or *Session 2*

⁹ To be maximally inclusive in identifying potential synaesthetes at this earliest stage, we repeated this process replacing 1SD with 1.5SD, and we repeated a third time where we compared colours by their colour category (i.e., we converted RGB values to the 11 basic colour terms in English, following Rinaldi, Smees, Alvarez, et al., 2019). Each method produced its own distribution of Letter-scores and Number-scores (from which we identified high-performing children 1.96SD over the mean; see below).

¹⁰ We point out for maximum clarity that the term '*high memory controls*' is used here for continuity with the literature, and does not suggest that children were assessed for their memory in any way *other than* by providing consistent colours for graphemes (within a single test session, while not being synaesthetes). Following the literature, we assume these non-synaesthetes performed well within the single session either by chance, by using a strategy, or by having a superior memory span (because they did not show the long-term consistency typical of a synaesthete, see below). The term for such children in the literature has been '*high-memory controls*' (e.g., Simner et al., 2009) which we continue here.

Number Score was significantly high for his/her age (i.e., $>1.96SD$ above the age-linked mean, as before).

In parallel, we also computed one final consistency score: the *Delayed consistency score*. This was an indication of which children had been consistent not within Session 1 or within 2, but across the 6-10 month interval separating the two sessions. For delayed consistency, we compared the first selection of colours in Session 1 with the first selection in Session 2 (e.g., the first of the three colours given for letter A in Session 1, compared to the first of the three colours given for letter A in Session 2). We computed Letter and Number Scores in the same manner as before, again using the Session 1 means to identify who was 1.96SD more consistent than the mean for his/her age. (This was a very conservative requirement, since it meant that true synaesthetes needed to be significantly more consistent across 6-10 months than their peers had been within the 10 minute test of Session 1.) Given all these measures, we divided our participants into three groups: *true synaesthetes* (consistent within Session 1, *and* consistent within Session 2, *and* consistent across sessions) versus *high-memory non-synaesthetes* (consistent in Session 1, but not in all three), versus *average-memory non-synaesthetes* (inconsistent in Session 1, and therefore not retested for synaesthesia).

Table 1.

Classification of children following screening for grapheme-colour synaesthesia after each Session. Ave-mem = *average-memory*; high-mem = *high-memory*. Shading indicates the age/gender breakdown for Session 1 categories (potential synaesthetes $n = 332$, *average-memory* controls $n = 663$). Note that ages shown are as of Session 1 although children in Session 2 were 6-10 months older.

			Age (in years)				
Status Session 1	Status Session 2	Gender	6	7	8	9	10
Potential synaesthete 332		F 168	29	42	38	36	23
		M 165	33	39	43	34	16
	Synaesthete 41	F 22	1	5	3	7	6
		M 19	1	5	5	7	1
	High-mem Control 261	F 137	25	35	33	29	15
		M 124	27	29	32	22	14

	Removed 30						
Ave-mem control 663		F 332	56	82	67	87	40
		M 331	69	74	82	74	32
	Ave-mem Control 605	F 318	52	77	64	87	38
		M 287	58	66	72	61	30
	Removed 58						

Note: Average memory controls were not retested in Session 2, but their numbers reduced in response to the removal of their matched potential synaesthete.

Identifying OLP synaesthetes

We identified OLP synaesthetes following Simner, Alvarez et al. (2019). This takes a similar approach to above, in that we first identified a group of potential synaesthetes who were consistent within Session 1 and we then used data from Session 2 to separate this group into *true synaesthetes* (who continued to be consistent in Session 2, and across sessions) and *high-memory non-synaesthetes* (who did not continue to be consistent after Session 1). Unlike above, children were given three consistency scores (each out of 26 letters) because they could show consistency (a) for *personality matches*, where gender is ignored (e.g., A is always friendly); (b) for *gender matches*, where personality is ignored (e.g., A is always female); and (c) for *strict matches*, where both personality and gender count (e.g., A is always a friendly female). Children identified as consistent within any of these scores were identified as *potential synaesthetes* from Session 1 (and subsequently re-classified after Session 2 as either *true synaesthetes* or *high-memory non-synaesthetes*).

As above, *high-memory non-synaesthetes* will have scored well in Session 1 either from memory alone or by having applied strategies that they failed to apply in subsequent testing (e.g., G is for girl therefore G is female). Recognising strategies is particularly important in this OLP test because responses are made from among only six choices (i.e., six faces), rather than the 23,050 colours in the grapheme-colour test. This means that chance-responding produces relatively consistent performance, so there is a risk of approaching ceiling if strategies are used even to a minor degree. For this reason (i.e., risk of strategies, small number of response-domains) we follow Simner, Alvarez et al. (2019) in determining consistency using a weighted scoring method, which scores rarer matches (e.g., F = male) more highly than common matches (e.g., F = female). We then applied

the thresholds from Simner, Alvarez et al. (2019) to identify children responding consistently for their age in any of their scores (at the 99th percentile from a Monte Carlo simulation of weighted scores: see Simner, Alvarez, et al., 2019).

As a result of these calculations, we identified 241 *potential OLP synaesthetes*, who had given consistent personifications for letters in Session 1. From among those who failed in Session 1, we identified 481 children to serve as *average-memory controls*. After Session 2, our *potential synaesthetes* were further divided into the categories of *true synaesthete* and *high-memory non-synaesthetes*, as shown in Table 2. Finally, we removed 127 children (82 *potential synaesthetes*, 45 *average-memory controls*) for not following task instructions (i.e. they choose the same gender for > 20/26 letters or the same personality for >16/26 letters; see Simner, Alvarez, et al., 2019) plus 157 controls who had been matched to potential synaesthetes who were themselves subsequently excluded; see Table 2.

Table 2.

Classification of children following screening for OLP synaesthesia after each Session. (Ave-mem = *average-memory*; high-mem = *high-memory*). Shading indicates the age/gender breakdown for Session 1 categories (*potential synaesthetes* $n = 241$, *average-memory controls* $n = 493$). Note that ages shown are as of Session 1, although children in Session 2 were 6-10 months older.

			Age (in years)				
Status Session 1	Status Session 2	Gender	6	7	8	9	10
Potential synaesthete 241		F 122	12	30	31	35	14
		M 119	22	30	25	29	13
	Synaesthete 41	F 20	0	3	7	6	4
		M 21	1	5	3	8	4
	High-mem Control 118	F 65	6	18	16	18	7
		M 53	8	12	13	16	4
	Removed 82						
Ave-mem control 493		F 248	36	46	64	70	32
		M 245	47	56	57	61	24
	Ave-mem Control 291	F 162	17	32	39	48	26
		M 129	14	24	34	42	15
	Removed 202						

Note: Average memory controls were not retested in Session 2, but their numbers reduced in response to the removal of their matched potential synaesthete.

What is the personality profile of a child synaesthete?

We next examined the personality traits of the different groups identified in our synaesthesia screening. These tests had identified 41 *grapheme-colour synaesthetes*, along with their 663 *average-memory controls* and 289 *high-memory controls*. The tests had also identified 41 *OLP synaesthetes*, along with their 291 *average-memory controls* and 118 *high-memory controls*. Since we did not anticipate a difference in our controls depending on which type of synaesthete we had allocated them to, we collapsed control groups to enlarge sample size. Hence our personality analyses will compare four groups: *grapheme-colour synaesthetes*, *OLP synaesthetes*, *average-memory controls* and *high-memory controls*.

Since our child-rated personality test was taken only by those aged 8 and over, it was taken by 30 *grapheme-colour synaesthetes* (15 female, 15 male, mean age = 9.16, SD = 0.83), 32 *OLP synaesthetes* (17 female, 15 male, mean age = 9.08, SD = 0.83), 209 *high-memory controls* (114 female, 95 male, mean age = 8.93, SD = 0.85) and 465 *average-memory controls* (243 female, 222 male, mean age = 8.98, SD = 0.85). In our parent-rated personality test, we had 278 parents. Of these, 15 were parents of *grapheme-colour synaesthetes* (11 female, 4 male, mean age = 8.40, SD = 1.16), 20 were parents of *OLP synaesthetes* (10 female, 10 male, mean age = 8.71, SD = 1.13), 114 were parents of *high-memory controls* (64 female, 50 male, mean age = 8.35, SD = 1.21) and 133 were parents of *average-memory controls* (64 female, 69 male, mean age = 8.32, SD = 1.18).

Since we will compare the personality traits of *grapheme-colour synaesthetes* and *OLP synaesthetes*, we therefore removed the five children who had both types of synaesthesia (*grapheme-colour and OLP*) because they could not be allocated to our mutually-exclusive groups (and we judged that $n = 5$ would be too small to explore personality within multiple-variant synaesthetes). The number of children in each group are shown in the analyses below, for child-rated personality and parent-rated personality respectively. We conducted multinomial log-linear regression analyses in R version 3.5.0 using the *nnet* package version 7.3-12 (Ripley & Venables, 2016). In this analysis we used personality scores as predictors, with membership in one of the four groups as the outcome. The reported changes in likelihood are in comparison to our control group,

treating our largest cohort as the reference group (i.e., *average-memory controls*; but see *Supplementary Information* (SI) at the end of the Chapter, for parallel models switching reference group to *high-memory controls*). We included age as a covariate given age-differences across groups ($F(4, 1086) = 3.93, p = .004$) and we followed standard approaches to ipsatize child-rated personality scores prior to our analyses, in order to control for the effect of acquiescence-bias in children (see Rinaldi et al., 2019).

Definitional BFI-44-C.

In our child self-rated questionnaire, we investigated differences between 25 *grapheme-colour synaesthetes*, 27 *OLP synaesthetes*, 405 *average-memory controls*, and 209 *high-memory controls*. Setting our reference group to *average-memory controls*, we found participants were significantly more likely to be synaesthetes if they had higher *Openness* scores, for both *grapheme-colour synaesthetes* and for *OLP synaesthetes* (see Table 2). Here, a one unit increase in *Openness* scores, which was rated on a five-point scale, corresponded to a 5.63 increase in odds ratio of being a *grapheme-colour synaesthete* (or a 463% increase in the odds), and a 4.22 increase in the odds ratio of being an *OLP synaesthete* (322% increase in odds). We next set our reference to *high-memory controls* and found a similar pattern (see Table 1, SI); an increase in *Openness* was associated with 3.87 increase in odds ratio (287%) in the relative odds of being a *grapheme-colour synaesthete* compared to a *high-memory controls*. There was also a 2.90 increase (190%) in the relative odds of being an *OLP synaesthete*, but this effect was only trending ($p = .078$; see Table 1, SI).

Table 3.

Group differences in child-rated personality using Multinomial Log-linear Regression with significant results shown in bold.

			Lower CI	Upper CI					%
		Co-	(Co-	(Co-		Wald	p-	Odds	Change
Group	Term	efficient	efficient)	efficient)	SE	z	value	Ratio	in Odds
Reference: Average memory controls									
High memory control	Intercept	0.12	-1.37	1.61	0.76	0.16	.874	1.12	12.85
	Neuroticism	0.07	-0.24	0.38	0.16	0.43	.699	1.07	7.02
	Openness	0.37	-0.06	0.81	0.22	1.68	.094	1.45	45.57
	Agreeableness	-0.06	-0.50	0.38	0.22	-0.26	.792	0.94	-5.71
	Conscientiousness	0.20	-0.21	0.61	0.21	0.97	.331	1.22	22.46
	Extraversion	0.01	-0.34	0.36	0.18	0.05	.960	1.01	0.88
	Age	-0.13	-0.31	0.05	0.09	-1.45	.146	0.88	-12.20
Grapheme-colour		-5.33	-9.30	-1.35	2.03	-2.63			
Synaesthete	Intercept						.009	0.00	-99.51
	Neuroticism	0.18	-0.57	0.92	0.38	0.46	.645	1.19	19.16
	Openness	1.73	0.45	3.00	0.65	2.66	.008*	5.63	462.68
	Agreeableness	0.61	-0.51	1.73	0.57	1.06	.288	1.84	83.83
	Conscientiousness	-0.37	-1.39	0.66	0.52	-0.70	.485	0.69	-30.59
	Extraversion	-0.29	-1.18	0.61	0.46	-0.63	.530	0.75	-24.92

Group	Term	Lower CI		Upper CI		Wald z	p-value	Odds Ratio	% Change in Odds
		Co-efficient	(Co-efficient)	(Co-efficient)	SE				
OLP Synaesthete	Age	0.15	-0.30	0.59	0.23	0.64	.522	1.16	15.72
	Intercept	-3.31	-6.97	0.35	1.87	-1.77	.076	0.04	-96.35
	Neuroticism	0.20	-0.55	0.94	0.38	0.52	.602	1.22	21.94
	Openness	1.44	0.28	2.59	0.59	2.44	.015*	4.22	321.62
	Agreeableness	-0.80	-1.80	0.20	0.51	-1.57	.116	0.45	-55.18
	Conscientiousness	0.72	-0.25	1.70	0.50	1.45	.146	2.06	106.24
	Extraversion	-0.37	-1.22	0.48	0.43	-0.86	.390	0.69	-31.06
	Age	-0.01	-0.44	0.41	0.22	-0.06	.955	0.99	-1.22

Note: * indicates significance at the $p = .05$ level. The model AIC = 1253.02, deviance = 1295.02.

Our data is summarized in Figure 1, which shows means scores for each group, for each of the five personality factors.

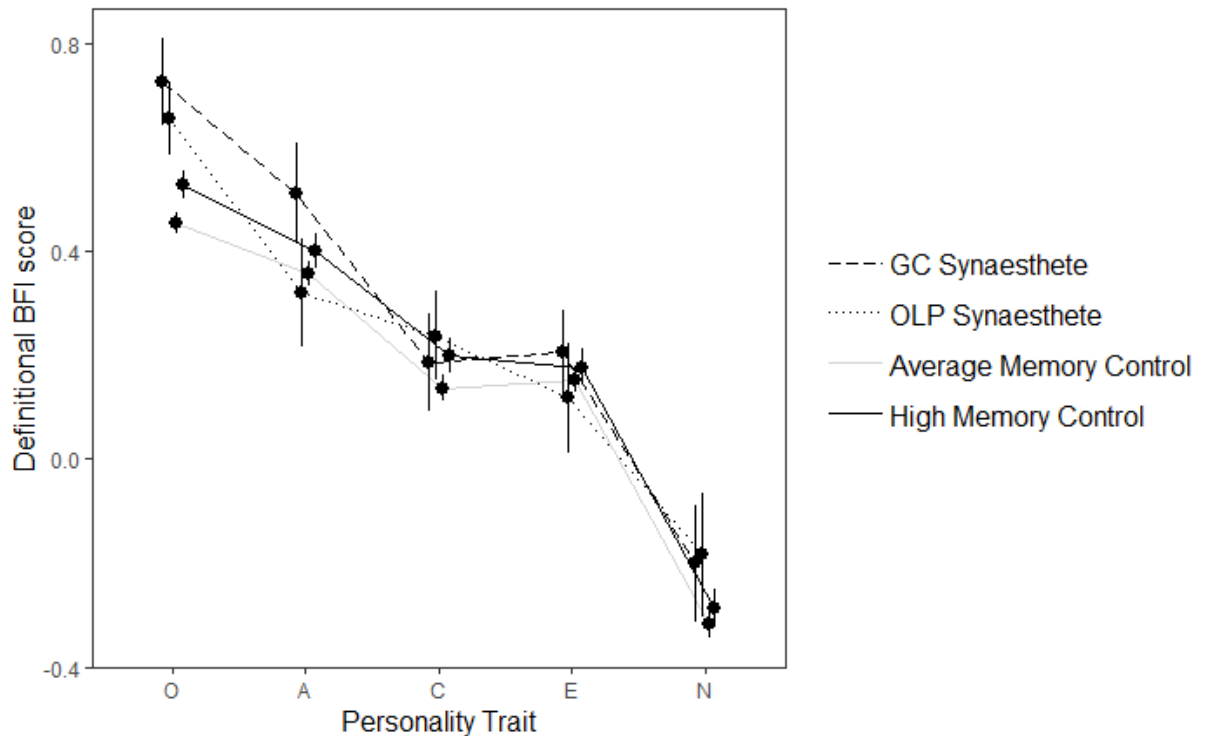


Figure 1. Means scores for each of the five personality factors in the (child-reported) Definitional BFI-44-C questionnaire, where O stands for Openness, A for Agreeableness, C for Conscientiousness, E for Extraversion and N for Neuroticism. Error bars show standard error of the mean. Note that dotted lines are synaesthetes and solid lines are controls.

BFI-44-Parent.

We next examined scores from our parent-rated questionnaire, based on 11 *grapheme-colour synaesthetes*, 16 *OLP synaesthetes*, 153 *average-memory controls*, and 114 *high-memory controls* whose parents had filled out the parent questionnaire. We began as before, by setting our reference group to *average-memory controls*, and again found evidence of a link between *Openness* and synaesthesia (see Table 3) – but this time only *grapheme-colour synaesthesia*. Higher *Openness* was significantly associated with being a grapheme-colour synaesthete, where a one unit increase in *Openness* scores, which again was completed on a five-point scale, gave an 8.18 increase (718%) in the relative odds of being synaesthetic compared to an *average-memory control*. We found a similar effect when setting our reference to *high-memory controls* (see Table 2, SI). Again, an

increase in *Openness* was associated with a significant increase in the relative odds of being a *grapheme-colour synaesthete* compared to a *high-memory control* (9.68 increase in odds or 868%). Parent-reports did not show the significant *Openness* link found earlier in child-reports for OLP synaesthetes, despite elevated odds at 129% in comparison to *average-memory controls*, and 170% in comparison to *high-memory controls*.

Our parent-rated data showed additional effects beyond those in the child-rated questionnaire, for two further traits. *Grapheme-colour synaesthetes* showed significantly lower *Extraversion* than *average-memory controls*, with a 68% reduction in odds of being synaesthetic for each unit of *Extraversion* (see Table 3). *Grapheme-colour synaesthetes* also showed lower *Extraversion* than *high-memory controls* (see Table 2, SI: when we set our reference as *high-memory controls*, *grapheme-colour synaesthetes* showed a 60% reduction in the odds of being synaesthetic for each unit increase in *Extraversion*). Finally, our parent-rated data showed that *OLP synaesthetes* were associated with higher *Conscientiousness* compared to *average-memory controls* (2.70 increase in odds or 170% change; see Table 3 below) but not compared to *high-memory controls* (see Table 2, SI).

Table 4.

Group differences in parent-rated personality using Multinomial Log-linear Regression with significant results shown in bold

		Lower CI		Upper CI				%	
		Co-	Co-	Co-		Wald		Odds	
Group	Term	efficient	efficient	efficient	SE	z	p-value	Ratio	Change
Reference: Average memory controls									
High memory control	Intercept	1.25	-2.27	4.77	1.80	0.70	.487	3.49	249.12
	Neuroticism	-0.02	-0.37	0.33	0.18	-0.09	.924	0.98	-1.67
	Openness	-0.17	-0.70	0.35	0.27	-0.63	.529	0.85	-15.45
	Agreeableness	-0.19	-0.58	0.21	0.20	-0.92	.358	0.83	-16.93
	Conscientiousness	0.38	-0.01	0.76	0.20	1.90	.057	1.46	45.58
	Extraversion	-0.23	-0.57	0.11	0.17	-1.31	.129	0.80	-20.48
	Age	-0.05	-0.26	0.16	0.11	-0.45	.656	0.95	-4.71
Grapheme-colour									
Synaesthete	Intercept	-6.67	-16.95	3.62	5.25	-1.27	.204	0.00	-99.87
	Neuroticism	-0.45	-1.31	0.41	0.44	-1.02	.307	0.64	-36.11
	Openness	2.10	0.21	3.99	0.96	2.18	.029*	8.18	718.49
	Agreeableness	-0.24	-1.23	0.74	0.50	-0.48	.632	0.79	-21.40
	Conscientiousness	0.33	-0.62	1.28	0.48	0.68	.495	1.39	39.08
	Extraversion	-1.14	-1.96	-0.32	0.42	-2.74	.006**	0.32	-68.08

Group	Term	Lower CI		Upper CI		Wald		Odds Ratio	% Change in Odds
		Co-efficient	Co-efficient	Co-efficient	SE	z	p-value		
OLP Synaesthete	Age	0.03	-0.50	0.57	0.27	0.13	.898	1.04	3.55
	Intercept	-8.11	-16.22	0.03	4.15	-1.95	.051	0.00	-99.97
	Neuroticism	0.18	-0.53	0.90	0.37	0.50	.617	1.20	20.06
	Openness	0.83	-0.44	2.10	0.65	1.28	.201	2.29	128.62
	Agreeableness	-0.70	-1.51	0.11	0.41	-1.69	.091	0.50	-50.20
	Conscientiousness	0.99	0.11	1.88	0.45	2.21	.027*	2.70	170.40
	Extraversion	-0.19	-0.89	0.51	0.36	-0.53	.593	0.83	-17.33
	Age	0.22	-0.20	0.65	0.22	1.02	.307	1.25	24.82

Note: * indicates significance at the $p = .05$ level. The model AIC = 567.54, deviance = 524.54

Our data is summarized in Figure 2, which shows mean scores for each group, in each of the five personality factors.

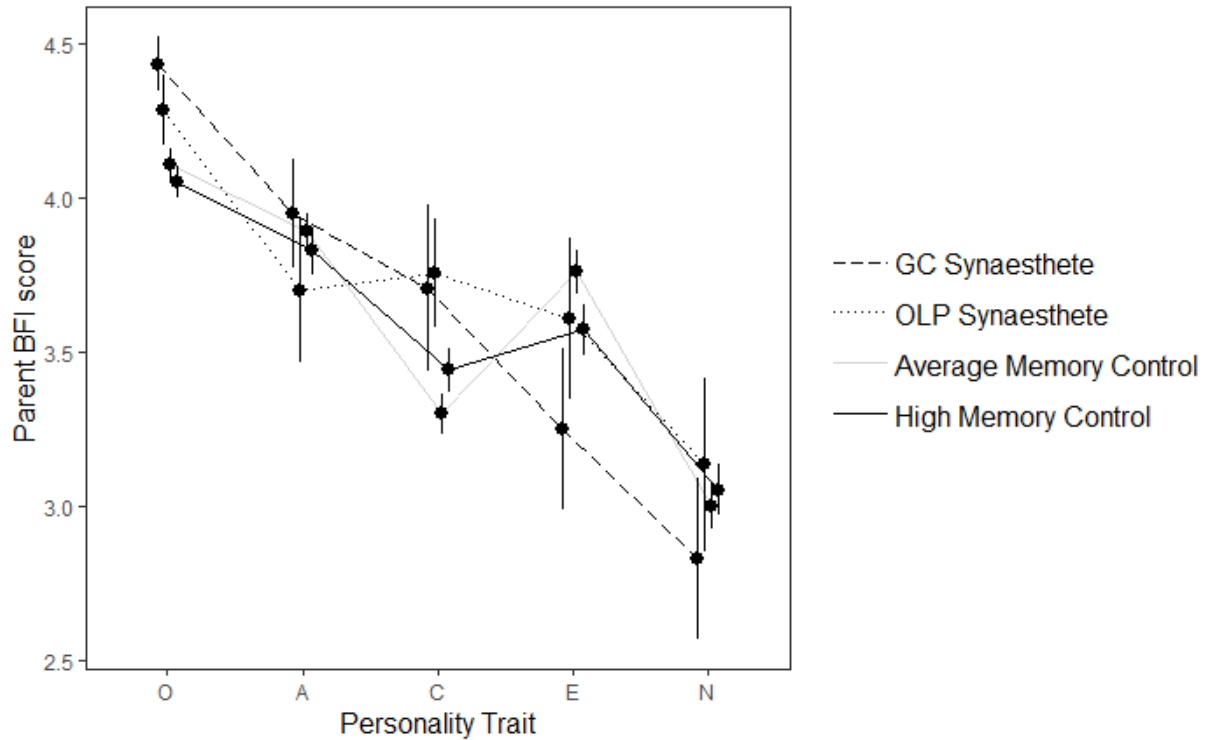


Figure 2. Means scores for each of the five personality factors in the (parent-reported) BFI-44-Parent questionnaire, where O stands for Openness, A for Agreeableness, C for Conscientiousness, E for Extraversion and N for Neuroticism. Error bars show standard error of the mean. Note that dotted lines are synaesthetes and solid lines are controls.

Discussion

In this study, we investigated whether child synaesthetes show a personality profile that sets them aside from their peers. We had several main aims with this research: firstly to extend previous personality findings to a randomly sampled group of verified synaesthetes; secondly, to extend these findings to children, during a period in development when synaesthesia is still emerging; and thirdly, to compare two different common subtypes of synaesthesia (grapheme-colour synaesthesia and OLP synaesthesia). We also included two types of non-synaesthete controls: *high-memory controls* (who can invent and recall synaesthesia-like associations in the short-term, but are not synaesthetes), and *average-memory controls* (who have average performance in this domain). These two groups allow us to estimate whether differences stem from cognitive factors such as

memory (in which case synaesthetes and high memories controls might score similarly), or whether they are tied to synaesthesia itself (in which case *synaesthetes* and *high-memory controls* would score differently).

Our principal finding was that synaesthesia, regardless of subtype, was associated with higher *Openness*, supporting the prediction that different variants of synaesthesia may share a unified personality profile. However, we also found type-dependent traits: grapheme-colour synaesthetes showed lower *Extraversion* compared to average and *high-memory controls*, while *OLP synaesthetes* showed higher *Conscientiousness* compared to *average-memory controls*. We discuss these findings in turn below.

Our finding that synaesthetes show higher *Openness* replicated important previous research by Banissy et al. (2013), Rouw and Scholte (2016) and Chun and Hupé (2016). All three had methodological differences to our own study, in which they had recruited synaesthetes and controls differently to each other, or mentioned synaesthesia during recruitment, or had not measured synaesthesia objectively. However, our results suggest their findings of high *Openness* were not due to methodological considerations, since we replicate this here with an unbiased sample of verified synaesthetes. We identified synaesthetes by objective measures, and by screening virtually the entire student bodies of 22 primary schools with almost no opt-outs (1%). Given this confidence, we might now ask why *Openness* is a trait found in synaesthesia, for both children and adults.

Openness is principally categorised by two main attributes; intelligence and creativity (Caspi et al., 2005). Since our synaesthetes scored higher in *Openness* compared to even *high-memory non-synaesthetes*, our finding is unlikely to be linked to intelligence alone. And indeed, there is independent evidence that both intelligence and creativity are elevated within synaesthesia. Synaesthetes not only score highly in intelligence-linked domains such as memory (Rothen et al., 2012), but also partake more often in creative activities and score higher in certain creativity tasks (Rothen & Meier, 2010b; Ward et al., 2008). Furthermore, Janik McErlean and Banissy (2016) found heightened *Openness to Experience* was not related to sensation seeking behaviour in adult grapheme-colour synaesthetes unlike in the general population (Garcia, Aluja, Garcia & Cuevas, 2005). This suggests, as do our findings, that synaesthesia is associated with a specific personality profile that may be distinct from non-synaesthetes. The fact that we have found synaesthesia-linked differences in *Openness* stemming back into childhood argues

against a model in which this trait develops over time by repeated exposure to synaesthetic sensations (e.g., repeatedly seeing colours enticing a synaesthete to paint; see Simner, 2019). The youngest children in our study are still in the process of developing their synaesthesia (see Simner, Rinaldi, et al., 2019) so would have had only nascent exposure to what will become lifelong associations. This suggests that other factors may be dictating personality profiles, and we return to this question further below, after reviewing our other key findings.

We also found two additional traits linked to synaesthesia, but each were tied to one particular variant of synaesthesia. Within parent-reported personality, grapheme-colour synaesthetes showed lower *Extraversion*. This effect has not been found in any of the three previous studies of grapheme-colour synaesthesia in adults (Banissy et al., 2013; Chun & Hupé, 2016; Rouw & Scholte, 2016; though not all tested grapheme-colour synaesthetes in isolation). Importantly, however, their earlier recruitment methods may have masked this effect because they relied on some degree of self-motivation from their participants (whilst there was no self-motivation required within our own sample). Put simply, any person willing to reach out to scientists, or willing to leave their contact details for future study, may be somewhat high on *Extraversion* already. This would be true of both synaesthetes and controls, meaning that matching recruitment across testing groups would not resolve this issue (i.e., selection is from people already *a priori* extraverted). An alternative explanation for our finding, however, is that we tested a group of children rather than adults, so it is possible that lower *Extraversion* pertains only to *young* synaesthetes. We have suggested this because one of the core elements of *Extraversion* is dominance, and this is known to increase from adolescence through to middle age (Caspi et al., 2005). It might be possible, therefore, that young synaesthetes had lower *Extraversion* simply because they have not yet developed in dominance. However, the fact that synaesthetes, only, showed this trait, suggests it is associated with childhood synaesthesia per se, rather than simply with childhood.

We additionally found higher *Conscientiousness* from parent-reports, comparing OLP synaesthetes to *average-memory controls*. This OLP-linked finding conflicts with Rouw and Scholte (2016), who found *decreased Conscientiousness* in their group of mixed synaesthetes. However, we noted earlier that Rouw and Scholte (2016) recognised synaesthetes by self-declaration alone, and that a suprisingly high number of self-declared ‘synaesthetes’ are not synaesthetes at all (Simner, Mulvenna, et al., 2006). Inattention is

a possible reason to incorrectly self-declare synaesthesia, and this trait is linked to low *Conscientiousness* (Grieve, 2012). A similar argument may explain why Rouw and Scholte found their self-declared synaesthetes to be high in *Neuroticism*, while our sample were not. *Neuroticism* is linked with hypochondria and pathologizing (Costa & McCrae, 1987) so might reasonably be high in a group of people incorrectly thinking they may have a rare neurodevelopmental condition. Nonetheless, it is also possible that our differences to Rouw and Scholte (2016) speak to age-differences in our samples: higher *Neuroticism* may evolve as synaesthetes age, perhaps as they recognise their differences, and/ or in parallel with other age-related increases in *Neuroticism* (Soto, John, Gosling, & Potter, 2011). However, the absence of high *Neuroticism* or low *Conscientiousness* in any other adult study of synaesthetes leads us to tentatively assume these effects may be related to self-declaration of synaesthesia and its known links to false-reporting.

Importantly, we found here that *Conscientiousness* was higher than in *average-memory controls*, not just for *OLP synaesthetes*, but potentially also for *high-memory non-synaesthetes*. (The comparison between *high-* and *average-memory controls* just missed significance at $p = .057$, and there was no difference between *high-memory controls* and *OLP synaesthetes*.) This is perhaps unsurprising given that some degree of conscientiousness is required to perform well in our diagnostic tests without synaesthesia. Specifically, many *high-memory controls* will have achieved their high OLP test-scores by applying strategies, or by trying hard to remember letter-face associations they gave earlier in testing -- both signs of high *Conscientiousness*. However, it is important to acknowledge a possible limitation in our study. Given the link between *Conscientiousness* and performing well in our diagnostic test (possibly by both *OLP synaesthetes* and *high-memory non-synaesthetes*), we tentatively suggest that *Conscientiousness* in synaesthetes may be a task-dependent confound, and we therefore take a conservative approach in giving this finding less weight than our other significant results (of higher *Openness* and lower *Extraversion*).

Finally, unlike Banissy et al. (2013), we found no indication that grapheme-colour synaesthetes were lower in *Agreeableness*. If low *Agreeableness* really were a trait tied to synaesthesia, this could logically arise as synaesthetes come to learn that they are different from their peers (i.e., leading to isolation and thereby low *Agreeableness*). Finding no similar effect in child synaesthetes is certainly consistent with this theory because personality traits arising from exposure to synaesthesia would logically be

limited in younger children (who have had less exposure). However, Banissy et al.'s *Agreeableness* finding was not replicated in the adult samples of either Chun and Hupé (2016) nor Rouw and Scholte (2016; although these latter did not focus solely on grapheme-colour synaesthesia). We simply note, therefore, that low *Agreeableness* has not been linked to grapheme-colour synaesthesia in children, nor has it been linked to synaesthesia more broadly in two out of three adult studies.

We end by considering the types of mechanisms that might lead to the personality profile we have identified here. One possible mechanism is via shared brain regions implicated in both personality and synaesthesia. It is interesting to note that both *Openness to Experience* and *Extraversion* (i.e., the key traits found here) share similar neurological underpinnings (Kennis, Rademaker, & Geuze, 2013). Both have been linked to networks that account for differences in sensitivity to reward (known as The Behavioural Approach System) and both traits are associated with overlapping brain activation in temporal and parietal regions, amongst others (See Kennis et al., 2013 for review). Additionally, both *Openness* and *Extraversion* have been linked to functional brain activation in similar areas to grapheme-colour synaesthesia (e.g., insula and dorsal prefrontal cortex; Kennis et al., 2013; Rouw, Scholte, & Colizoli, 2011). And there is similar overlap in structural terms: both *Extraversion* and grapheme-colour synaesthesia have been linked to cortical differences in volume and surface area in the fusiform gyrus and superior temporal gyrus (Riccelli, Toschi, Nigro, Terracciano, & Passamonti, 2017; Rouw et al., 2011). Additionally, both *Openness* and grapheme-colour synaesthesia have been linked to differences in cortical thickness and surface area of the anterior cingulate gyrus, inferior parietal cortex and lateral occipital gyrus (Riccelli et al., 2017; Rouw et al., 2011). Shared regions are therefore important for both synaesthesia and *Openness/Extraversion*, suggesting that personality differences may emerge from these shared neurological roots. Of course, we must acknowledge the possible circularity in this account. Regions associated with synaesthesia (i.e., regions found when scanning synaesthetes) may be nothing more than personality differences themselves. This is especially true for *structural* imaging studies, which do not elicit synaesthesia during scanning, and might therefore have highlighted differences between synaesthetes and controls which were personality determined.

In conclusion, we have tested a large sample of child synaesthetes, avoiding recruitment bias and other testing confounds as far as was possible. We have found that child

synaesthetes do indeed have personality differences compared to their peers. We have found that children with either grapheme-colour synaesthesia or OLP synaesthesia are higher than their peers in *Openness to Experience* (replicating previous findings in adult synaesthetes). We have also found that, compared to *average-memory controls*, child *grapheme-colour synaesthetes* are lower in *Extraversion*, while child *OLP synaesthetes* are higher in *Conscientiousness* (although we conservatively link this latter with the possibility of task demands). With respect to previous findings shown in adult synaesthetes but not found here, we point to one of two interpretations: aging effects (perhaps for low *Agreeableness* and/or high *Neuroticism*), or methodological issues in earlier studies (perhaps for high *Neuroticism* and/ or low *Conscientiousness*). Finally, we note that differences might also arise from random variability in relatively small sample sizes, given the rareness of this fascinating condition.

Chapter 3: Supplementary Information

Additional models switching the reference to high-memory controls

Child-rated Definitional BFI-46-C.

Table 1 below shows an increase in *Openness* was associated with 3.87 increase (287%) in the relative odds of being a grapheme-colour synaesthete compared to a *high-memory controls*. There was also a 2.90 increase (190%) in the relative odds of being an OLP synaesthetes, but this effect was only trending.

Table 1.

Group differences in child-rated personality using Multinomial Log-linear Regression with significant results shown in bold

Group	Term	Co-efficients	Lower CI (Co-efficient)	Upper CI (Co-efficient)	SE	Wald z	p-value	Odds Ratios	% Change in Odds
Reference: <i>High memory controls</i>									
Average memory controls	Average	-0.12	-1.61	1.37	0.76	-0.16			
	Intercept						.873	0.88	-11.45
	Neuroticism	-0.07	-0.38	0.24	0.16	-0.43	.699	0.93	-6.56
	Openness	-0.37	-0.81	0.06	0.22	-1.68	.094	0.69	-31.26
	Agreeableness	0.06	-0.38	0.50	0.22	0.26	.792	1.06	6.06
	Conscientiousness	-0.20	-0.61	0.21	0.21	-0.97	.332	0.82	-18.34
	Extraversion	-0.01	-0.36	0.34	0.18	-0.05	.960	0.99	-0.88
	Age	0.13	-0.05	0.31	0.09	1.45	.146	1.14	13.90

Group	Term	Co- efficients	Lower CI (Co- efficient)	Upper CI (Co- efficient)	SE	Wald z	p- value	Odds Ratios	% Change in Odds
Grapheme- colour Synaesthete	Intercept	-5.45	-9.51	-1.38	2.08	-2.62			
	Neuroticism	0.11	-0.66	0.87	0.39	0.28	.009	0.00	-99.57
	Openness	1.35	0.05	2.65	0.66	2.04	.783	1.11	11.35
	Agreeableness	0.67	-0.48	1.82	0.59	1.14	.042*	3.87	286.61
	Conscientiousness	-0.57	-1.62	0.49	0.54	-1.05	.256	1.95	94.92
	Extraversion	-0.30	-1.21	0.62	0.47	-0.63	.292	0.57	-43.31
	Age	-0.30	-1.21	0.62	0.47	-0.63	.528	0.74	-25.57
	Age	0.28	-0.18	0.74	0.23	1.18	.238	1.32	31.81
OLP synaesthete	Intercept	-3.43	-7.19	0.33	1.92	-1.79			
	Neuroticism	0.13	-0.64	0.90	0.39	0.33	.073	0.03	-96.78
	Openness	1.06	-0.12	2.25	0.60	1.76	.739	1.14	13.94
	Agreeableness	-0.74	-1.78	0.29	0.53	-1.41	.078	2.90	189.84
	Conscientiousness	-0.74	-1.78	0.29	0.53	-1.41	.159	0.48	-52.46
	Extraversion	0.52	-0.48	1.53	0.51	1.02	.310	1.68	68.43
	Extraversion	-0.38	-1.25	0.49	0.45	-0.86	.392	0.68	-31.66
	Age	0.12	-0.32	0.56	0.22	0.53	.597	1.13	12.54

Note: * indicates significance at the $p = .05$ level. The model AIC = 1295.02, deviance = 1253.02.

Parent-rated BFI-44-C.

Table 2 below shows an increase in *Openness* was associated with a significant increase in the relative odds of being a grapheme-colour synaesthetes compared to a *high-memory controls* (9.68 increase in odds ratio or 868%). Grapheme-colour synaesthetes additionally showed a 60% reduction in the odds of being synaesthetic for each unit increase in *Extraversion*.

Table 2.

Group differences in parent-rated personality using Multinomial Log-linear Regression with significant results shown in bold

Group	Term	Co-efficient	Lower CI (Co-efficient)	Upper CI (Co-efficient)	SE	Wald z	p-value	Odds Ratios	% Change in Odds
Reference: High memory controls									
Average memory controls	Intercept	-1.25	-4.77	2.27	1.80	-0.69	.487	2.87	-71.33
	Neuroticism	0.02	-0.33	0.37	0.18	0.09	.925	1.02	1.69
	Openness	0.17	-0.35	0.69	0.27	0.63	.529	1.18	18.26
	Agreeableness	0.19	-0.21	0.58	0.20	0.92	.358	1.20	20.37
	Conscientiousness	-0.38	-0.76	0.01	0.20	-1.90	.057	0.69	-31.31
	Extraversion	0.23	-0.11	0.57	0.17	1.31	.189	1.26	25.75
	Age	0.05	-0.16	0.26	0.11	0.45	.656	1.05	4.94
Grapheme-colour Synaesthete	Intercept	-7.92	-18.19	2.35	5.24	-1.51	.131	3.63	-99.96
	Neuroticism	-0.43	-1.29	0.43	0.44	-0.98	.326	0.65	-35.02
	Openness	2.27	0.38	4.16	0.96	2.36	.018*	9.68	868.43
	Agreeableness	-0.06	-1.04	0.93	0.50	-0.11	.912	0.95	-5.39

Group	Term	Co-efficient	Lower CI (Co-efficient)	Upper CI (Co-efficient)	SE	Wald z	p-value	Odds Ratios	% Change in Odds
OLP synaesthete	Conscientiousness	-0.05	-1.00	0.91	0.49	-0.09	.925	0.96	-4.46
	Extraversion	-0.91	-1.73	-0.10	0.41	-2.20	.028*	0.40	-59.87
	Age	0.08	-0.45	0.62	0.27	0.31	.760	1.07	8.66
	Intercept	-9.35	-17.49	-1.21	4.15	-2.25	.024	8.71	-99.99
	Neuroticism	0.20	-0.52	0.92	0.37	0.55	.585	1.22	22.09
	Openness	0.99	-0.27	2.26	0.65	1.54	.123	2.70	170.43
	Agreeableness	-0.51	-1.32	0.30	0.41	-1.24	.214	0.60	-40.06
	Conscientiousness	0.62	-0.26	1.50	0.45	1.38	.168	1.86	85.77
	Extraversion	0.04	-0.65	0.73	0.35	0.11	.913	1.04	3.96
	Age	0.27	-0.16	0.70	0.22	1.24	.216	1.31	30.99

Note: * indicates significance at the $p = .05$ level. The model AIC = 566.54, deviance = 524.5

Chapter 4

Do the colors of educational number-tools improve children's mathematics and numerosity?

Chapter Summary

In Chapters 2 and 3 I explored personality differences to answer the question of what makes child synaesthetes different from non-synaesthetes aside from their synaesthesia. In these earlier chapters I examined personality, and here I turn to cognition – specifically number cognition. I first explore this in non-synaesthetes who have ‘synaesthesia-like’ experiences. The current chapter takes a coloured number tool used widely in primary schools in the UK and tests whether some children spontaneously internalize its colour associations, and if so, how this effects their numerical abilities.

This study capitalized on the fact that we were planning to screen more than 3000 children for grapheme-colour synaesthesia. Although we would primarily focus on the 2% identified as synaesthetes, we would also have data from the remaining 98% (i.e., colour associations for letters and numbers). We therefore formulated hypotheses based on what we might expect for these children, given their multisensory environment. Within this environment were coloured number-tools, which allowed us to formulate hypotheses for how these might impact on their learning. This chapter is the outcome, and has been accepted for publication as Rinaldi, L.J., Smees, R, Alvarez, J, Carmichael, D. C., & Simner, J (*in press*). Do the colors of educational number-tools improve children's mathematics and numerosity? *Child Development*. We note here that this chapter uses American English spellings (e.g., color) because it is in press in an American journal and where additional models were included in a supplementary information in our article submission, here they have been instead provided at the end of the chapter.

Abstract

This study examined how colored educational tools improve children's numerosity ('number sense') and/or mathematics. We tested children 6-10 years ($n=3236$) who had been exposed to colored numbers from the educational tools Numicon (Oxford University Press, 2018) or Numberjacks (Ellis, 2006), which map colors to magnitudes or Arabic numerals respectively. In a free-association task pairing numbers with colors, a subset of children spontaneously provided colors matching these schema. These children, who had internalized Numicon (colored magnitude), showed significantly better numerosity but not mathematics compared to peers. There was no similar benefit from internalizing Numberjacks (colored numerals). These data support a model in which colored number-tools provide benefits at different levels of numerical cognition, according to their different levels of cross-modal mappings.

Introduction

Early-years educators often use educational aids in mathematics, and these tools provide physical representations to make numbers more concrete (see Wing & Tacon, 2007). A large proportion of these tools pair numbers with colors, and these colored number-aids are aimed particularly at school children 4-11 years. For example, in one commonly used tool, *Numicon* (Oxford University Press, 2018), the numbers one to ten are physically represented as colored shapes with differing numbers of holes to represent magnitude (See Figure 1). The pairing of number with color and shape is assumed to promote mental imagery and this visualization of numbers is seen as key to the learning approach (see Wing & Tacon, 2007).

Tools such as *Numicon* are widely used in primary school education in the UK (Day & Lockwood, 2008; Devon Primary Math Team, 2006; Ewan & Mair, 2002; Wing & Tacon, 2007) as well as across Europe and in countries worldwide. The feature of interest in the current study is the color of these tools, since each number has an associated color which is consistent across all sets. Here, we investigated the degree to which the colors of tools such as *Numicon* may help children internalize numbers, and how this might impact on different numerical cognition skills. We look particularly at mathematics and numerosity ('number sense'; see below) and present a model predicting the efficacy of different colored number tools, in which colors bind to either magnitude or Arabic numerals and thereby influence different levels of numerical processing. We test our theory with data from over three thousand children who have been exposed to *Numicon* (encoding colored magnitude) or to a second tool which pairs colors to Arabic numerals (see below). We begin with a brief overview of the scientific literature on mathematical educational aids, and then introduce our model, hypotheses, and study.



Figure 1. Numicon Shapes: A graphic representation of the *Numicon* shapes one to ten, which are individual 3-D plastic forms with the colors and configurations shown here.

Numicon represents just one example of the larger class of “math manipulatives”, an umbrella term for objects used in mathematics to help make abstract numerical concepts more concrete (Clements & McMillen, 1996). These objects include not just *Numicon* shapes but also cubes, number rods, counting posters, and so on. Evidence suggests the use of these math manipulatives is a good pedagogical technique for teaching numeracy. For example, a meta-analysis by Carbonneau, Marley and Selig (2013) examined 55 studies comparing over 7000 students across the schooling years 6-18 years. Results showed that students using math manipulatives performed better than students using abstract symbols alone. Results were particularly striking in some areas over others, for example, with fractions showing a greater effect size than algebra or arithmetic. This suggests that math manipulatives can aid in different aspects of number cognition although the reasons for this are not entirely clear. Carbonneau et al. (2013) found the size of the effect was better when there was more emphasis on instruction given by educators, and was also influenced by age, with younger children showing moderate effects and older children showing smaller effects. However, this meta-analysis did not include *Numicon*, making it unclear how this particular math manipulative might fare.

Scientific validation for *Numicon* in particular (which we will use in our testing for the current study) has been attempted from a series of studies, many of which show numerical trends of improvements in mathematics for children using these tools, but often without statistical validation or control conditions (e.g., Education Leeds, 2008; Ewan and Mair, 2002; Tacon, Atkinson, & Wing, 2004, but see Nye, Buckley, & Bird, 2005). But the

strongest case for support of *Numicon* was a randomized control trial by the *Education Development Trust* (Churches, 2016). This looked at two different math interventions including *Numicon*, within a sample of 875 low-performing students in School Years 1, 2 or 3 (between the ages 5-10 years). Approximately half the children were assigned to the *Numicon* group, and the rest were assigned to a control group where teachers continued teaching as they had before the study. Mathematical ability was measured before and after study using the *Progress in Math* tests which cover the UK mathematics curriculum (Clause-May, Vappula, & Ruddock, 2004). Churches (2016) found that *Numicon* was the only statistically successful intervention. In a replication one year later, controls and intervention children were assigned *within* each school to eliminate *a priori* differences across schools. Children in School Year 2 were randomly assigned to control or *Numicon* intervention groups and there was again a moderate but significant effect of improvement in the intervention group compared to controls.

A recent trend in the UK along with many other countries (e.g., USA) has been for more evidence-based policy (Gorard, See, & Siddiqui, 2017) and to the best of our knowledge, the study by Churches (2016) is the only randomized control trial on *Numicon*. The current study aims to contribute to the evidence-base on color-coded tools such as *Numicon* by taking a novel approach in investigating how and why *Numicon* might aid in number cognition, and placing these findings within a general theory accounting for the benefits of math manipulatives. We examined one specific aspect of *Numicon* in particular – its colors – to attempt to understand which features of *Numicon* might aid in which aspects of numerical processing and why. Tools such as *Numicon* have been colored deliberately on the assumption this plays a role in their educational benefits, so it is important to examine whether this choice has a meaningful effect. In our study, we took whole schools which have already been using *Numicon* and we looked across its pupils to find those who had internalized the *Numicon* colors. For this we ran a pre-test asking children to simply free-associate colors to numbers. We then measured how many times their color associations married with *Numicon* colors (e.g., the number five is red in *Numicon*; did they free-associate 5 = red?). Comparing to chance levels, we took this as an index of whether *Numicon* colors had been mentally internalized by each individual child, and then used this metric to divide children into two groups: those who had internalized *Numicon* colors and those who had not. Finally, we took independent tests of numerical cognition across groups. If children had integrated the colors of *Numicon* into

their mental number system, we asked whether, and in what areas, they might perform better in numerical cognition. Our theory predicts that improvements would be tied to the nature of the cross-modal coloring expressed by the manipulative itself (i.e., whether the tool associates colors with magnitudes or numerals; see below).

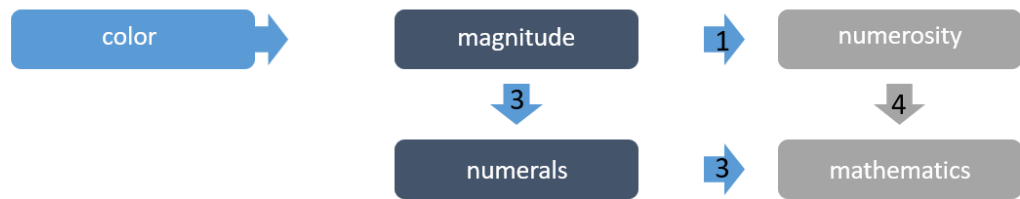
To understand our theory better, consider that we gave two types of numerical tests: a test of mathematics and a test of numerosity. Numerosity is our intuitive “number sense” which allows us to understand magnitudes without knowing the exact amount. This sense of numerosity relies on an *approximate number system* (ANS) which comprises a set of mental processes that approximately encode magnitudes (Dehaene, 2001). One common way of measuring numerosity is to ask participants to quickly discriminate between two arrays of dots, such as an array of black dots next to an array of white dots. Although the dots may be displayed too briefly to count, it is still possible to intuit whether the black or white dots were more numerous. Adults are able to do this with great success (Barth, Kanwisher, & Spelke, 2003) and can discriminate dot arrays which differ by a factor of 1.15 or more (e.g., Lipton & Spelke, 2004). Even infants show evidence of an early ANS, and although their ANS is initially imprecise, it develops over time (Feigenson, Dehaene, & Spelke, 2004; Xu & Spelke, 2000). Importantly, the cross-modal nature of *Numicon* pairs color with – specifically – magnitude (rather than numerals): its plastic shapes have pierced holes to represent magnitudes from one to ten and do not resemble numerals (see Figure 1). For this reason, we theorize that any advantage from the cross-modal influence of color would correspond to better performance in numerosity in particular.

We also included a comparison condition, which is a source that again matches numbers with colors, but this time pairs colors to Arabic numerals (rather than magnitudes per se). This baseline comes from a BBC television show (*Numberjacks*; Ellis, 2006) widely viewed by primary school children in the UK in which animated colored numerals 0 to 9 solve mathematical problems. The show was first released in the UK but has since been syndicated to countries worldwide, including the USA. This baseline allows us to test whether associating colors with numerals is beneficial in itself for processing magnitude, in which case a child who had internalized either *Numicon* or *Numberjacks* colors should show an advantage in numerosity. Alternatively, colored numerals may fail to benefit numerosity per se, because there is no cross-modal coding of color to magnitude itself.

Our model accounts for cross-modal advantages via the known benefits of “Dual-coding” (e.g., Clark & Paivio, 1991; Paivio, 1969). Here, colors would give number an enhanced level of encoding through a greater number of memory cues. These additional memory cues are assumed to strengthen representations and in turn facilitate retrieval. In the case of *Numicon*, these additional cues are bound at the level of magnitude. In contrast, *Numberjacks* colors are bound at the level of Arabic numerals (but not magnitude directly), and this leads to our first prediction: Children who have internalized *Numicon* colors (dual-coding magnitude) should correspondingly have better performance in a test of numerosity, but these benefits would not be seen in children who have internalized *Numberjacks* (dual-coding numerals). Our second prediction is that children who have internalized *Numberjacks* colors might have better performance in our mathematics test, because this was designed around the UK curriculum and many of its questions are phrased using numerals.

Our third prediction comes from a consideration of how numerosity and mathematics interact. Importantly, although numerosity and mathematics ability are related (i.e., adults and children with high numerosity perform better in mathematics; Anobile, Stievano, & Burr, 2013; Chen & Li, 2014; Halberda Mozzocco & Feigenson, 2008), this relationship is assumed to be directional. Wong, Ho, and Tang (2016) have suggested a direction of causality in that a better ANS aids the ability to map numerosities to numerals and consequently improve math ability. This directional mapping from numerosity to mathematics suggests that children who have internalized *Numicon* colors (i.e., dual-coded magnitude) may perform better not only in numerosity but also in tests of mathematics. However, children who have internalized *Numberjacks* colors (i.e., dual-coded numerals) may not see similar benefits in numerosity. Our fourth prediction is that the known empirical relationship between numerosity and mathematics (improved numerosity correlates with improved mathematics; e.g., Anobile, Stievano, & Burr, 2013) is itself unrelated to color and will therefore operate irrespective of whether children have internalized colors from any device. Our model, and its four predictions are represented in Figure 2 below.

Numicon



Numberjacks

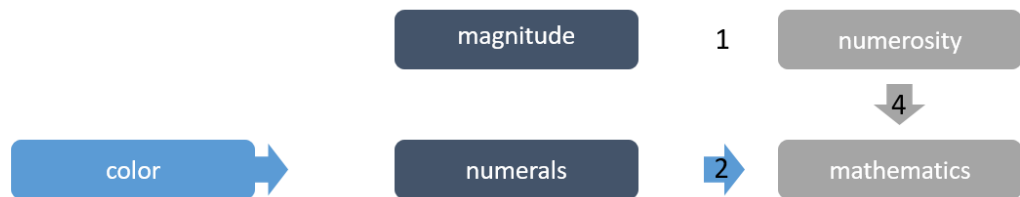


Figure 2. Modelling Math manipulatives: Left column shows the dual coding of color at different levels of representation from two types of math manipulatives (*Numicon* and *Numberjacks*). Middle column shows color is mapped directly to magnitude in *Numicon* but to numerals in *Numberjacks*. Final column shows the testing measures where our model predicts effects. Blue arrows show hypothesized dual-coding benefits from color, and grey arrows represent benefits unrelated to color. Our four hypotheses (see text) are mapped onto the model as numerals 1-4.

There have been very few studies of how numerosity (or indeed mathematics) might be improved by colored number tools like *Numicon* or *Numberjacks*. But in addition to the one study reviewed above (showing the efficacy of *Numicon* in mathematics), there is reason to think math manipulatives might well have an impact on numerosity. DeWind and Brannon (2012) showed that numerosity can indeed be improved with intervention: they trained 20 adults across six sessions on numerosity judgements with accuracy feedback and found that numerosity improved significantly. This suggests that the ANS is malleable so might be influenced by tools. Another line of evidence, this time relating to color in particular, comes from color-number associations in unusual populations. *Grapheme-color synesthesia* occurs in approximately 1.5% of adults (Simner et al., 2006; Carmichael, Down, Shillcock, Eagleman, & Simner, 2015) and children (Simner et al., 2009) and causes lifelong, automatic, quasi-idiosyncratic associations between colors and numerals (or between colors and letters/ words). There is a growing body of evidence that synesthetes perform better in certain cognitive domains (e.g., memory for words; see

Meier & Rothen, 2013b) and this has been linked by some to the same benefits of dual-coding we explore here (Radvansky, Gibson, & McNerney, 2011); hence synesthetes may have enhanced cognition because they dual-code graphemes with color information. We ask here, therefore, whether similar mechanisms of dual coding can also enhance cognition in non-synesthetes (and we return to this comparison with synesthesia in the *Discussion*).

Finally, regarding *Numberjacks*, there is some evidence that children do benefit from watching educational television: a longitudinal study by Wright et al. (2001) suggested preschool children had better receptive vocabulary, number skills and engaged in more reading if they had watched child-audience informative programs at ages 2-3 years. Furthermore, a meta-analysis of 24 studies in 15 countries suggested that children who watched the child-oriented show *Sesame Street* (Ganz Cooney & Morrisett, 1969-2019) performed better across basic literary, numeracy, science, health and safety, and pro-social reasoning (Mares & Pan, 2013). Together these results suggest that math manipulatives can influence learning, that benefits may come from colored numbers, that children can learn from television shows, and that both mathematics and numerosity show improvements from intervention. Finally, we point out that an alternative prediction is that colors might produce a *negative* effect on children's learning by increasing the cognitive load on mathematical thinking (e.g., McNeil, Uttal, Jarvin, & Sternberg, 2009). If colors become a distraction to learning they may inhibit numerical processing, or might directly inhibit processing certain types of math functions over others (e.g., inhibit arithmetic, where multiple colors could compete).

In summary, we present a study in which we elicited free-associations between colors and numbers from a group of over three thousand children. We used their responses to divide children into groups: those that had internalized *Numicon* colors versus those who had not; and those who had internalized *Numberjacks* colors versus those who had not. Finally, we tested whether children with internalized colors were better in tests of numerosity and mathematics. Our approach differs to previous studies in that we do not compare children according to whether or not they *use* tools such as *Numicon* (e.g., Churches, 2016), but we instead take a cohort who *all use these tools* and look instead at whether or not they have internalized the colors. Our key prediction is that those who had internalized *Numicon* (pairing color with magnitude) but not *Numberjacks* (pairing color with numerals) should correspondingly have better performance in a test of numerosity (i.e.,

magnitude judgements). A second prediction is that internalizing the colors of *Numberjacks* might be associated with increased mathematics performance by its direct color-coding of numerals. A third prediction is that internalizing the colors of *Numicon* may perhaps be associated with improved mathematics if any benefits from colored magnitude feed forwards into mathematics. A final prediction is that a relationship between numerosity and mathematics is also likely to exist independently of whether children have internalized colors.

Methods

Participants

In our Numerosity assessment we tested 3236 children aged 6-10 years (mean age = 7.95; SD = 1.22). These were 1571 girls (mean age = 7.95, SD = 1.22) and 1665 boys (mean age = 7.95, SD = 1.22). Of these children, 92.5% were native English speakers. Children were recruited from 22 Infant and Primary schools across East and West Sussex in the south of England (n = 15 from East Sussex, n = 7 from West Sussex) and were in School Years 2-5 (for ages see Table 1). As an indicator of affluence/poverty (Taylor, 2018) the mean school-level free school meal (FSM) percentage was 13.44 %, where the national average from the same year is 14.5%, and our schools ranged in FSM status from 0.7% to 38.1%. In our Curriculum Math assessment, we tested a sub-group of these children, comprising n=2519 (mean age = 8.40; SD = 0.97; 1228 girls, mean age = 8.39, SD = 0.97; 1291 boys, mean age = 8.40, SD = 0.96) who were children in School Years 3-5 only (see below for why Year 2 were tested for numerosity, but not curriculum math).

We also tested but excluded an additional 63 children. Of these, 20 were removed because they did not complete the tasks, and a further 33 experienced a technical error. Nine were flagged by teachers at our request as being newly arrived in the UK with particularly low levels of English, and one final child was out of year group (i.e., her chronological age did not match the rest of her class). Our study was approved by the local university ethics board and testing took place from October 2016 to the end of April 2017.

Materials and Procedure

Children were tested in class groups of approximately 30 within their classrooms, and they completed up to three tasks in the following order: a Curriculum Math test, a Numerosity test, and Colored numbers test. School Years 3-5 completed all tests, while Year 2 completed the latter two only because they had not yet covered enough of the math curriculum to be tested on mathematics (see below). Together our tests typically lasted for 5-10 minutes each but were interspersed with other activities (e.g., personality testing) to be reported elsewhere. These activities separated our tests by approximately 20 minutes.

Curriculum Math Test

we developed a short math test for children in School Years 3-5 based on the UK Primary school curriculum (“The national curriculum in England: Key stages 1 and 2 framework document,” 2013). This test was presented on paper and there were 47 questions in total, one question for each of the 7-9 sub-sections of the math curriculum per years 1-6 (see example questions in Figure 3). For each child however, our test assessed knowledge of the curriculum for the child’s current school year and two years prior. For example, Year 3 students start the test with questions from the Year 1 curriculum. (Since there is no set UK math curriculum prior to Year 1, students in Year 2 could not complete an equivalent test and were therefore excluded from mathematics testing). Children were given five minutes to complete as much as they could, as quickly and accurately as possible, and were not expected to go beyond their current year (e.g., Year 3 pupils start with Year 1 questions and, in general, are not expected to get to Year 4 questions). Children who got further than their current year were marked for all correct questions. Teachers and researchers gave no help to children, except with reading if necessary, and no feedback was given.

Please write your answers in the boxes

1. Fill in the missing number below

30	40	50	60	70		90
----	----	----	----	----	--	----

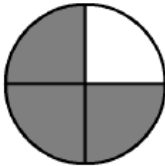
2. Please subtract

$$\begin{array}{r} 17 \\ - 8 \\ \hline \end{array}$$

3. Please divide

 $10 \div 2 =$

4. What fraction is shown in grey? Write the fraction in the box



=

Figure 3. Mathematics Testing Materials. Example questions from our math test based on Year 2 (age 6-7) curriculum content.

Numerosity Task

The numerosity task (and the colored numbers task which follows) was presented on electronic tablets. Children were each given a touch screen Acer Aspire SW3-016 or Acer One 10 tablet, which ran on Intel® Atom TM x5-Z8300 Processors. These ran on Windows 10 and had 10.1" LED backlight touchscreens (1280 x 800 pixels).

Children were given the Panamath Numerosity Dot Task (Halberda et al., 2008) which required them to make a judgement based on dots on the screen. Children saw a cluster of white dots on the left side of the screen, and a cluster of black dots on the right side of the screen. Their task was to press one of two buttons (marked with a white or black sticker) to indicate whether there were more white dots or black dots. We used the default Panamath settings (Halberda et al., 2008) which generate an adjusted level of difficulty

based on each child's age. The length of time for this task is adjustable and we set the task to run for two minutes. Children were told they would play a game in which they saw black and white dots on the screen. They were instructed to press a button to show whether there were more black or white dots. They were told they would not have time to count the dots so they should make their best guess as quickly as possible.

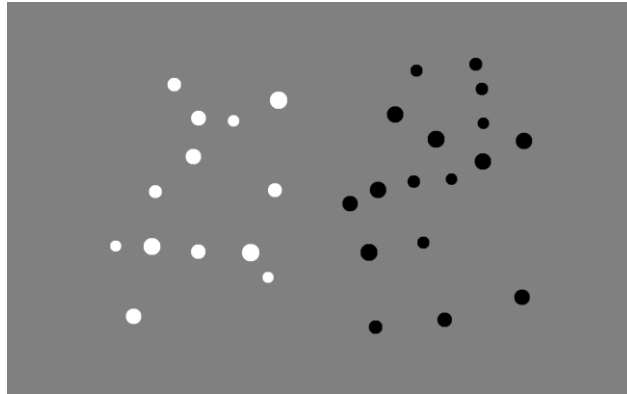


Figure 4. Numerosity Task. A screenshot of the numerosity dot task. Here the correct answer is ‘black’ (i.e., screen shows 13 white dots and 17 black dots so black dots are more numerous).

Free Association: Number-Color Pairing Task.

We also tested children on a free association number-coloring task developed by our lab. In this test, children saw the numbers 0-9, individually in a random order, and were asked to think of the ‘best’ color for each number. Children chose their color using an on-screen color picker which appeared on the right-hand side of the screen and consisted of a vertical bar which could be dragged up and down to select hue¹¹. To the right of the hue bar was a 10x10 grid of color-swatches which allowed children to also select the exact shade. Within this 10x10 grid, luminance varied along one axis and saturation along the other, with axes randomly alternating from trial to trial. For example, if the child saw “5” and wanted to select a certain shade of red, (s)he would first drag the hue bar down to red, and would then inspect the 10x10 shade box to find the exact luminance and saturation of red required. Manipulating the color picker provided children with a choice from 25,600 discrete colors in RGB (red, green, blue) color space. Children were first trained

¹¹ A screenshot of this test was not included in the journal submissions due to publication costs, but can be seen in Appendix D.

to use the color-picker, which they managed without difficulty. Each trial began with a random initial setting. Graphemes appeared in lowercase black font, 2.5cm high, in a typeface suitable for children (Sassoon Infant®). Children inspected the number, which appeared on the left of the screen, and then made a color choice from the right-hand palette. Once the color choice was made, the program advanced to the next trial. (This task was also used to test an unrelated set of hypotheses to be reported elsewhere which required letters interspersed with the numbers, and presented three repeated blocks. For the present study, however, we report only the colors of numbers, taken from the first block only.)

Results

Participant exclusions

In our analyses, we examine two color schemas (*Numberjacks*, *Numicon*) with different exclusion criteria. For our *Numberjacks* analyses, there were no exclusions; i.e., all children were included since all were likely to have seen this extremely popular television show. For *Numicon*, we included only children whose teachers incorporated *Numicon* into their current teaching programs (since we could not otherwise guarantee children had seen these school-specific products). We therefore excluded 395 (non-*Numicon*) participants from our Curriculum Math analysis (this left $n = 2124$ Years 3-5; mean age = 8.39; SD = 0.96) and the same 395 participants from our Numerosity analysis (3 of whom had already been excluded earlier for technical failures in Numerosity, see *Participants*; this left $n = 2844$ Years 2-5; mean age = 7.89; SD = 1.23). (We point out that the same number of participants were excluded from both tests, despite more children taking Numerosity than Curriculum Math overall. This is simply because there were no exclusions among the extra (Year 2) children taking Numerosity). The year group and gender of participants within each analysis is shown in Table 1.

Table 1

Number of participants broken down by analysis, year group, and gender

Year group	<i>Numberjacks</i>		<i>Numicon</i>	
	Female	Male	Female	Male
2	343	374	343	374
3	421	435	361	362
4	392	416	332	366
5	415	440	341	365
Total	1571	1665	1377	1467

Data preparation

In order to compare children's color choices with those from *Numicon* and *Numberjacks* we first coded children's color-choices into color categories (red, green, blue, yellow etc.). We next compared their chosen colors to those found within the comparison schemas of *Numicon* and *Numberjacks*. Details of this coding procedure are given below.

Color Categorization Coding.

For our color-categorizing, we took the RGB color space co-ordinates from each child's numbers 0-9 and transformed these into the 11 basic color categories of English (black, white, red, blue, yellow, green, orange, pink, purple, grey, brown; Berlin & Kay, 1991) using the following method. We based our categorizations on the XKCD color survey (Munroe, 2010) in which color co-ordinates within RGB color space were named with color-labels by 222,500 participants. We aimed to use XKCD color survey data to establish the boundaries in color space for each of the 11 basic color categories in English (e.g., what is the boundary of the color red? What is the boundary of the color blue? etc.). Once done, we would use these boundaries to classify children's color co-ordinates into the 11 basic categories of English.

The participants of the XKCD color survey data (Munroe, 2010) named color co-ordinates using 949 color-terms (e.g., red, burgundy, pea green) so we first sorted these verbal color-labels into their 11 basic color categories. For this we were able to use definitions from the Oxford English Dictionary (OED) to classify 474 of these terms into either one color category (e.g., "navy" = blue), two categories (e.g., "violet" = blue + purple) or three categories (e.g., "violet pink" = blue + purple + pink). For the remaining

475 color-labels which had no clear OED definition, we recruited six researchers of color and sensory processing who were naive to the hypotheses of the experiment, to serve as coders. These coders were shown each color patch from Munroe (2010; which he had subsequently condensed by finding the central RGB of each repeated color-label using a stochastic hillclimbing algorithm). Coders were shown these patches on-screen alongside the names of 11 basic color categories. Coders were asked to simply select the best color category for each, and to select up to two categories where necessary. Coders agreed on all but 18 colors, and for these, both color categories were included (e.g., disagreement between blue and black resulted in both categories being accepted) which produced up to three color categories per item. This method provided us with boundaries in color space for each color category (red, green, blue etc.) which we could now apply to our children's RGB data. The outcome of our coding was that each child's color choice was now categorized within the 11 basic color terms of English.

Matching Colors to Schema.

Next we counted how many times each child had chosen a color for a number that matched either the *Numicon* or the *Numberjacks* schemas (see Table 2 below for the color-categories of these schemas which were rated by two independent coders with 100% agreement). For example, if the child had chosen red for the number 5, this would count as a match for *Numicon* (whose 5 is red) but not for *Numberjacks* (whose 5 is blue). In cases where children's colors had been categorized as more than one color (e.g., a certain shade of turquoise was blue + green) a match was counted if either of the colors matched with the given schema. Each child received a single score for each schema (i.e., a *Numicon* score and a *Numberjacks* score) which was the total number of matches out of a maximum of nine for *Numicon* (1-9) and out of ten for *Numberjacks* (0-9).

We next established how many matches would constitute chance levels, using a Monte Carlo approach which simulated 10,000 children making free-associations between colors and numbers. Specifically, the simulation began with the 11 colors in English, which were *a priori* weighted to reflect how often they were chosen by children across our entire data set (e.g., blue was chosen more frequently than orange so was weighted accordingly). These weighted colors were then selected at random (with replacement) in ordered sets of nine (for *Numicon*) or sets of 10 (for *Numberjacks*). We repeated this 10,000 times and

compared how each set matched to *Numicon* colors (or *Numberjacks* colors). In our *Numicon* simulation, for example, if the first color in the set matched the color of the *Numicon* shape for 1, this was a “match”. If the second matched the color of the *Numicon* shape for 2, this is was another “match”. This gave a match-score out of 9 for *Numicon* – and we did this repeatedly for 10,000 repetitions. This simply allowed us to establish the probability of matching to these schema by chance. Based on the conventional alpha of $p < .05$ we found the minimum number of matches to *Numicon* needed to exceed chance levels was five (and five matches was significant at $p = .011$). Five matches was also the appropriate statistical cut-off for *Numberjacks* (where five matches was significant again at $p = .011$). Given these analyses, we categorized children as using a *Numicon* or *Numberjacks* color-scheme if they had five or more matches (to *Numicon* or *Numberjacks* respectively), while children with four or fewer matches were considered to *not* be using these schema.

(A reviewer has asked us to also include an alternative approach where we identified internalizers against pure chance by running an equivalent Monte Carlo analysis but without weighting colors to reflect how often they were chosen by children; see *Supplementary Information* (SI) at the end of the chapter. Either method identifies internalizers (at 5 matches as shown above, or 4 matches as shown in *SI*) and will produce exactly the same pattern of results in our subsequent analyses below. See SI Tables 3-6 for parallel analyses.)

Table 2

Color associations for numbers 0-9 in *Numicon* and in *Numberjacks*.

Number	<i>Numicon</i> color	<i>Numberjacks</i> color
0	n/a	green
1	orange	purple
2	blue	orange
3	yellow	pink
4	green	blue
5	red	blue
6	blue	yellow
7	pink	red

8	green	blue
9	purple	green

Within the total sample of children who had been exposed to *Numicon* ($n = 2844$; Years 2-5), we found 26 children (0.9%) had internalized *Numicon*'s colors (i.e., they used *Numicon* as a coloring strategy more often than chance would predict). And within the total sample for *Numberjacks* ($n = 3236$; i.e., all children; Years 2-5) we found 100 (3.1%) had internalized *Numberjacks* colors. Within School Years 3-5 only (i.e., the cohort for our Mathematics testing) these numbers were 1% (21 out of 2124) and 3.1% (78 out of 2524) respectively.

There was no overlap between *Numicon*-internalizers and *Numberjacks*-internalizers (as expected, since these use different color-schemes).

Modelling the influence of color schemas on Numerosity and Math ability.

Here we test the hypotheses that children who internalize *Numicon* colors may have better numerosity or curriculum math abilities than those who do not. Children's binary classification of using or not using the *Numicon* strategy to color numbers was entered into two hierarchical regression models to compare their performance first on the numerosity test and then on the curriculum math test. Our analyses will allow us to determine whether children who free associate the colors of *Numicon* are better in a test of numerosity and/or a test of mathematics (and we will then do similarly for *Numberjacks* colors).

Numicon in Numerosity. Our dependent measure, percent correct in numerosity, had a negatively skewed non-normal distribution. This was due in part to the nature of the score we used (percent correct, 50% being chance), and in part due to participants performing well on the task, so we took a bootstrapping approach in our regression model. (We did not use an alternative output from this test, a Weber fraction, because the Weber fraction cannot produce a score for children at or around chance level which is a valid score in our analysis.) Along with *Numicon* strategy (using or not using) we included chronological age as a predictor in step one because our data suggest that older children were significantly more likely to internalize *Numicon* colors than younger children ($\chi^2(4) =$

10.62, $p = .03$). Both predictors had a significant effect on numerosity: older children, and those who had internalized *Numicon* colors, had better numerosity scores (see Table 3). The relationship between *Numicon* and numerosity equated to a gain of around 5% in numerosity scores for children internalizing *Numicon* colors. In order to further aid interpretation of this effect for *Numicon* colors, we investigated the Hedges g (quasi-equivalent to Cohen's d but for unequal groups). This effect size was small to moderate, Hedges $g = 0.34$. However, since this includes the influence of age, we re-examined Hedges g within a single age group (age 9, because this contained the largest set of *Numicon* internalizers) and produced a Hedges $g = 0.37$, suggesting the effect of *Numicon* is small-to-moderate.

Table 3

Numicon as a predictor of numerosity ability with 95% confidence intervals in brackets. Figures are based on 1000 bootstrap samples. Chronological age was entered as years in decimals.

	B	SE B	β	p
Step 1				
Constant	72.40 (69.14 – 75.70)	1.65		.001
Age	1.66 (1.28 – 2.02)	0.19	.17	.001
Step 2				
Constant	72.46 (69.23 – 75.76)	1.65		.001
Age	1.65 (1.27 – 2.00)	0.19	.17	.001
<i>Numicon</i> Integration	3.10 (-0.33 – 5.87)	1.50	.03	.038

Note: $R^2 = .03$ for step 1; $R^2 = .03$ for step 2, R^2 change = .001

Numicon in Curriculum Math. Each correct answer on the math test was given a score of 1 and these were summed to generate an overall mark. These were converted to z-scores to allow us to compare scores across years, given that children in different years saw different questions. We entered our z-scores as the dependent measure in our regression model, along with age and *Numicon* strategy as predictors (with *Numicon* strategy as a dummy variable: 'using strategy' = 1 and 'not using strategy' = 0). Age was centered around the mean chronological age of the year-group, because each child received a test

appropriate to his or her school year but could be older or younger *within the year*. Although age itself was a significant predictor of math ability ($\beta = 0.25$, $p < .001$), the *Numicon* strategy (using or not using) was not ($\beta = .02$, $p = .465$; see Table 4).

To explore this null result, we performed a Bayes factor analysis to determine whether we have enough evidence to accept the null hypothesis (Dienes, 2014). Our Bayes analysis assumes a half-normal distribution (Dienes, 2014) and we took our informative prior (i.e., a previous study against which to gauge our own findings) from a study showing improvement in math from a *Numicon* intervention (Education Leeds, 2008). This chosen prior (unlike, say, Churches, 2016) provided the statistical information necessary to calculate a Bayes Factor (i.e., a mean difference between groups that can be standardized, and the standard error of this mean). Bayes factors lie on a continuum in which scores less than 0.33 constitute evidence for the null hypothesis, and scores above 3 indicate evidence for the experimental hypothesis (Dienes, 2014). Our moderate Bayes Factor (BF = 0.15) was indeed less than 0.33, allowing us to accept the null hypothesis (Dienes, 2014) with sufficient power to conclude there is no difference in math performance between children who had or had not internalized the *Numicon* color-system.

Table 4

Numicon as a predictor of mathematics ability with 95% confidence intervals in brackets. Chronological age is mean centered, within each school year.

	<i>B</i>	SE <i>B</i>	β	<i>p</i>
Step 1				
Constant	0.02	0.02		.451
Age	0.73	0.06	.25	<.001
Step 2				
Constant	0.02 (-0.02 – 0.06)	0.02		.412
Age	0.73 (0.61 – 0.85)	0.06	.25	<.001
<i>Numicon</i> Integration	-1.55(-0.57– 0.26)	2.12	-.02	.465

Note: $R^2 = .06$ for step 1; $R^2 = .06$ for step 2

We end this section by pointing out that our pattern of results remains identical, even if we re-inserted all excluded children from our *Numicon* analyses. (These children had been excluded because they were not using *Numicon* in their current class, but were

nonetheless likely to have been exposed to *Numicon* in younger years, given the schools we tested.) Figures 1-6 in *Appendix F* show histograms illustrating the number of matches to each schema for these participants, as well as our corresponding analyses re-inserting excluded children; our pattern of results remain the same (See *SI*: Tables 1 and 2).

Numberjacks in Numerosity. We turn now to the color-schema of *Numberjacks*. We performed the same regression analysis on our Numerosity data as above, but this time using the binary classification of whether children did, or did not color numbers according to *Numberjacks*. Our results show that age was again a significant predictor of numerosity performance, but *Numberjacks* was not (see Table 5). We again ran a Bayes factor, here using an uninformed prior (in the absence of a suitable *Numberjacks* study for comparison) within Rouder and Morey's (2012) Bayes factor calculator for regression models (found at <http://pcl.missouri.edu/bayesfactor>). Our JZS Bayes Factor was 0.05 which again is under .33 lending strong support for the null hypothesis (M. Lee & Wagenmakers, 2014).

Table 5

Numberjacks as a predictor of numerosity ability with 95% confidence intervals in brackets. Figures are based on 1000 bootstrap samples. Chronological was entered as years in decimals.

	<i>B</i>	SE <i>B</i>	β	<i>p</i>
Step 1				
Constant	72.99 (69.81 – 76.28)	1.61		.001
Age	1.57 (1.20 – 1.92)	0.18	.16	.001
Step 2				
Constant	72.98 (69.80 – 76.28)	1.61		.001
Age	1.57 (1.21 – 1.92)	0.18	.16	.001
<i>Numberjacks</i> Integration	0.12 (-2.50 – 2.34)	1.23	.002	.936

Note: $R^2 = .02$ for step 1; $R^2 = .02$ for step 2

Numberjacks in Curriculum Math. Finally, we repeated our analysis investigating whether the *Numberjacks* strategy (used or not used) in coloring numbers predicted math

ability. Our results showed that age was again a significant predictor for math but *Numberjacks* was not significant (see Table 6), and a moderate-to-strong Bayes factor of 0.09 confirmed our strong support for the null hypothesis.

Table 6

Numberjacks as a predictor of mathematics ability with 95% confidence intervals in brackets. Chronological age is mean centered, within each school year.

	<i>B</i>	SE <i>B</i>	β	<i>p</i>
Step 1				
Constant	-0.001	0.02		.968
Age	0.72	0.06	.24	<.001
Step 2				
Constant	0.001 (-0.04 - 0.04)	0.02		.952
Age	0.72 (0.61 - 0.83)	0.06	.24	<.001
<i>Numberjacks</i> Integration	-0.01 (-0.21 - 0.23)	0.11	.002	.905

Note: $R^2 = .06$ for step 1; $R^2 = .06$ for step 2

Relationships between Numerosity and Math ability.

It is important to note that numerosity skills usually correlate with math ability (Halberda et al., 2008), and for this reason we verified whether these also correlated within our own cohort. As expected, there was a significant relationship between numerosity scores and mathematics scores, such that children scoring highly on one measure were likely to score highly on the other ($r = .26$, $p < .001$). In other words, although internalizing *Numicon* colors predicted numerosity and did not predict math ability, there was nonetheless a significant relationship between numerosity and math (as we would expect, whether or not children had internalized colors).

Discussion

Our paper set out to investigate how colored numbering within the educational devices *Numicon* and *Numberjacks* might aid children's numerical cognition. We did this by

identifying whether or not children had internalized the colors from each device, and the extent to which this aided them in a curriculum math test, and in a test of numerosity. To do this, we first asked children to free-associate colors to numbers and we inspected their responses to determine whether they had followed the schema of either *Numicon* or *Numberjacks* colors. We tested the math and numerosity skills of children who had internalized these colors, and compared them to controls who had not internalized either schema. Our model first predicted that *Numicon*, but not *Numberjacks*, would correspond to better performance in numerosity (i.e., sense of magnitude) because *Numicon* maps colors directly to magnitudes, while *Numberjacks* maps only to numerals. Our data supported this prediction: children who had internalized the colors of *Numicon* performed significantly better in numerosity.

The second prediction from our model was that internalizing *Numberjacks* (colored numerals) might bring benefits in curriculum mathematics because many of our questions were numeral-based. This prediction was not supported. Relatedly we hypothesized that *Numicon* colors (which aid magnitude processing) might have a “knock on” effect in the same curriculum math test. Again this third prediction was not supported. We therefore conclude that within manipulatives such as *Numberjacks* which link color to numerals, the color itself does not aid in math ability. And that although manipulatives linking color to magnitude (*Numicon*) are associated with benefits in numerosity, the dual-coding of color to magnitude does not easily transfer to Arabic symbols – or may not help in mathematics even if it does so. Instead, our fourth hypothesis was supported: there was a significant correlation between numerosity and mathematics (see also Halberda et al., 2008) but this was not influenced by the colors of math manipulatives. We represent these findings in our updated model, shown in Figure 5. Although our model is phrased in terms of two particular math manipulatives (*Numicon* and *Numberjacks*), it makes generalizable predictions beyond these exemplars, and extends to any math manipulatives using color in its approach to teaching math.

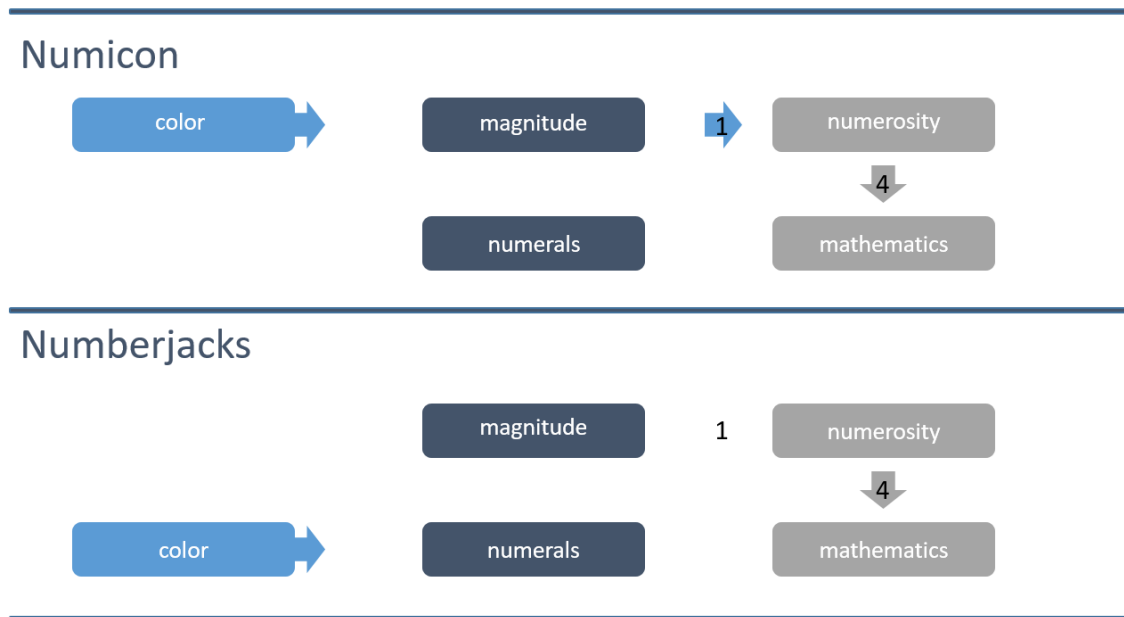


Figure 5. A Revised Model of Math manipulatives: Left column shows the dual coding of color at different levels of representation from two types of math manipulatives (*Numicon* and *Numberjacks*). Middle column shows color is mapped directly to magnitude in *Numicon* but to numerals in *Numberjacks*. Final column shows the testing measures where the model predicts effects. Blue arrows show hypothesized dual-coding benefits from color, and grey arrows represent benefits unrelated to color. Our data supported hypotheses 1 and 4.

Overall, our data suggests that an advantage in numerosity may come from *Numicon*'s colors. We have attributed this benefit in numerosity to *color* in particular because this was our independent variable (i.e., our groups were divided on whether they showed evidence of having internalized colors). Hence we have followed the standard empirical approach of attributing significance to the feature that was manipulated. However, it is logically possible of course (as in any study) that color may influence numerosity only indirectly, via some other correlating feature within *Numicon* (e.g., it could be that the shape of *Numicon* aids numerosity, and children who notice shape also happen to notice color). But there is no evidence in our study of any 'middle-man' influence, so we follow the assumptions of Occam's razor in attributing advantages in numerosity to the internalization of *Numicon*'s colors, in particular.

So why might colored magnitude aid numerosity, but colored numerals *not* aid mathematics? And is this finding to be expected? A test-case for the impact of colors on numerical processing might be to re-examine whether benefits in math are seen in

grapheme-color synesthetes. We saw earlier that synesthetes' dual-coding of color with graphemes improves cognition (e.g., memory for words) but evidence within numerical cognition has been somewhat equivocal: Green and Goswami (2008; see Simner and Bain, 2018 for in depth analysis) found that three out of eight synesthetic children with colored numerals showed superiority in mathematics, with a group trend $p = .09$. But their recruitment methods could have encouraged high performing children irrespective of synesthesia (see Simner & Bain, 2018 for discussion). We are therefore in the process of administering a mathematics (and numerosity) test to approximately 40 synesthetic children whom we have recruited using random sampling methods. In summary, studies to date suggest synesthetic dual-coding of numerals may not bring unambiguous benefits in mathematics testing – which mirrors our findings here – and our future studies are exploring this further.

We should of course acknowledge the small number of children within our sample who will have had synesthesia – and even color vision deficiencies – although these will have made only a very small contribution to our study given our large sample. For example, given the low prevalence of grapheme-color synesthesia in the population (e.g., Simner et al., 2009) 99% of our sample will not have synesthesia, although some nonetheless demonstrated memory associations linking colors and numbers, as we have shown. Learned associations such as these can sometimes be difficult to distinguish from genuine synesthesia in a behavioral sense (e.g., Meier & Rothen, 2009) but they have different neurological correlates. Elias, Saucier, Hardie, and Sarty (2003) compared genuine number-color synesthesia against a case where colored numbers had been acquired from the environment (by a lifetime of cross-stitching, in which threads are colored and numbered). Although both cases performed similarly in behavioral testing, only synesthesia resulted in activation of the dorsal visual stream when manipulating numbers, suggesting synesthetes alone possess the quasi-perceptual phenomena that is unique to synesthesia. The case of acquired colors from cross-stitch needles is directly equivalent to our own cases here, suggesting that the children in our study who had internalized colors would likely be using similar, non-synesthetic neurological mechanisms.

It is important to clarify our claim that the colors from these math tools (*Numicon* and *Numberjacks*) were ‘internalized’ by some children. Our criterion was that children had

to free-associate to the *Numicon* (or *Numberjacks*) colors more often than chance would predict. We assume that exceeding chance means that some psychological strategy was used, and this is the basis of our assumption that colors were ‘internalized’. We point out that our ‘internalizing’ threshold of five or more matches to the nine colors of *Numicon* may seem small by intuition alone, but statistically-speaking this is highly improbable. And perhaps most importantly, our data show that ‘internalizers’ were indeed a meaningful cohort, because they were also the category who performed better in numerosity; i.e., this categorization had a detectable impact on scores. For this reason we are confident that our samples were meaningfully divided into children who have, or have not, internalized colors from being exposed to colored number-tools.

Although we found no effect of either *Numicon* or *Numberjacks* in our curriculum math test, it is important to point out this does not mean *Numicon* and *Numberjacks* do not improve math. Indeed prior studies testing *Numicon* would suggest otherwise (e.g., Churches, 2016). Here we can conclude only that math improvement in earlier studies is unlikely to stem from *Numicon*’s colors. It may therefore be the shape qualities of *Numicon* which improve math, or indeed some interaction between color and shape. And it is important to point out any limitations of our findings. We have assumed that *Numicon* colors improve sense of magnitude but the reverse might also be true: children with better numerosity ability may be better able to integrate colors into their magnitude schema. The nature of regression statistics do not allow us to infer the direction of causality, although findings elsewhere suggest that exposure to colored numbers in *Numicon* does causally induce changes in numerical cognition (Churches, 2016). We therefore tentatively assume in the absence of direct counter-evidence that the dual-coding of color aids in magnitude estimation rather than vice versa.

We point out that ours is the first study to examine the impact of *Numberjacks* on mathematical literacy and our results point to *Numberjacks* colors being influential at a surface level (i.e., children do internalize its colors) but not at a conceptual level (this did not lead to improvements in numerosity or math). One consideration, however, is that *Numicon* is actively taught at school, while *Numberjacks* is watched passively at home. Carbonneau et al. (2013) found that math manipulatives have an increased effect on children’s learning when there is more emphasis on instruction given by the educator. We might therefore have found increased impact of colored numerals if these were used actively in the classroom, and we are now categorising schools according to their colored

numeral displays (e.g., wall posters) in order to assess the impact this might have on math attainment.

We point out that only a small amount of variance was captured by our significant model (involving *Numicon* colors and numerosity) and this equated to a small-to-moderate Hedges g of 0.37. However, this effect size must be taken in context, and is almost certainly because we took an extremely indirect measurement of whether *Numicon* and *Numberjacks* had been internalized: we did not ask children for *Numicon*/ *Numberjacks* colors, and we did not mention *Numicon*/ *Numberjacks* to them in any way. We simply asked children to color numbers in any way they wished, but would likely have found a clearer influence of *Numicon* on numerosity if we had instructed children to recall *Numicon* colors directly. (We avoided this because we did not want children to think about math manipulatives in the context of our math/numerosity tests.) Nonetheless, even with our highly indirect measure, the interaction between *Numicon* and numerosity was significant and equated to a gain of around 5% in numerosity scores for children internalizing *Numicon* colors. Overall this suggests that internalising colored magnitudes might indeed aid in numerosity ability in a way that is important to acknowledge.

In conclusion, we found that some children internalize number-color associations from the educational tools *Numicon* or *Numberjacks*, which pair colors with magnitudes or Arabic numerals respectively. In the former case, we found a significant improvement in children's numerosity abilities, and conclude that *Numicon*'s iconic representation of magnitude may help its colors become integrated into the ANS as a proxy for quantity. We found no benefits in mathematics testing, and no benefits in either numerosity or math for the colored numerals of *Numberjacks*. Together, our findings suggest that teaching magnitude-color patterns in education may be beneficial for the ANS in children's developing number cognition. Our findings would be of interest to a wide audience, including educationalists or researchers of developmental numerical cognition, or researchers of multisensory integration in learning, or indeed visual psychophysicists (we introduced a novel psychophysical metric for color categorisation). Finally, given that math manipulatives are common interventions for children with disabilities, our findings might also be relevant to clinical practitioners, and indeed to anyone interested in the benefits of internalizing environmental color. In summary, our results speak to the theoretical boundaries of multisensory learning, and to a fascinating interplay between numbers, colors, and education.

Chapter 4: Supplementary Information

Models including the excluded sample

Numicon in Numerosity

Table 1

Numicon as a predictor of numerosity ability with 95% confidence intervals in brackets. Figures are based on 1000 bootstrap samples. Chronological was entered as years in decimals.

	<i>B</i>	SE B	β	<i>p</i>
Step 1				
Constant	72.98 (69.89 – 76.14)	1.58		.001
Age	1.57 (1.22 – 1.93)	0.18	.16	.001
Step 2				
Constant	73.06 (69.93 – 76.19)	1.58		.001
Age	1.56 (1.21 – 1.93)	0.18	.15	.001
<i>Numicon</i>	3.22 (-0.69 – 5.81)	1.33	.03	.016
Integration				

Note: $R^2 = .02$ for step 1; $R^2 = .03$ for step 2; R^2 change = .001

Numicon in Maths

Table 2

Numicon as a predictor of mathematics ability with 95% confidence intervals in brackets. Chronological age is mean centered, within each school year.

	<i>B</i>	SE B	β	<i>p</i>
Step 1				
Constant	-0.01(-0.04– 0.04)	0.02		.968
Age	0.72 (0.61 – 0.83)	0.06	.24	<.001
Step 2				
Constant	0.01 (-0.04 – 0.04)	0.02		.976
Age	0.72 (0.61 – 0.84)	0.06	.24	<.001
<i>Numicon</i>	-1.27(-0.50 – 0.24)	0.19	-.01	.499
Integration				

Note: $R^2 = .06$ for step 1; $R^2 = .06$ for step 2

Pure Chance Monte Carlo analysis and resulting analyses

As in our main text we ran a Monte Carlo simulation, but with the following minor change. In place of weighting colours by how often they were chosen by the children in our sample

we did not use any weighting and entered the 11 colors in English equally. Then as before colors were then selected at random (with replacement) in ordered sets of nine (for *Numicon*) or sets of 10 (for *Numberjacks*). We repeated this 10,000 times and compared how each set matched to *Numicon* colors (or *Numberjacks* colors). Based on the conventional alpha of $p < .05$ we found the minimum number of matches to *Numicon* needed to exceed chance levels was 4 for both *Numicon* and *Numberjacks* at $p = .006$ and $p = .009$ respectively. Using this threshold we categorized children as using a *Numicon* or *Numberjacks* color-scheme if they had four or more matches (to *Numicon* or *Numberjacks* respectively), while children with three or fewer matches were considered to *not* be using these schema. This method identifies 206 children (7.2%) as having internalized *Numicon*'s colors, and 335 (10.4%) as having internalized *Numberjacks* colors. Within School Years 3-5 only (i.e., the cohort for our Mathematics testing) these numbers were 7.8% (166 out of 2124) and 9.8% (247 out of 2524) respectively. Below we present each of our analyses reported in the main manuscript again but using this less conservative threshold.

Numicon in Numerosity. As before we included chronological age as a predictor in step one because our data suggest that older children were significantly more likely to internalize *Numicon* colors than younger children ($\chi^2(4) = 10.62, p = .03$). Both predictors had a significant effect on numerosity: older children, and those who had internalized *Numicon* colors had better numerosity scores. See Table 3. This effect was small Hedges $g = 0.16$.

Table 3

Numicon as a predictor of numerosity ability. Figures are based on 1000 bootstrap samples. Chronological was entered as years in decimals.

	<i>B</i>	SE <i>B</i>	β	<i>p</i>
Step 1				
Constant	72.99	1.67		.001
Age	1.57	0.19	.16	.001
Step 2				
Constant	73.00	1.67		.001
Age	1.56	0.19	.15	.001

<i>Numicon</i>	1.49	0.66	.03	.025
<i>Integration</i>				

Note: $R^2 = .02$ for step 1; $R^2 = .03$ for step 2 R^2 change = .001

Numicon in Curriculum Math. As in the manuscript we entered our z score math as the dependent measure in our regression model, along with age and *Numicon* strategy as predictors. As in the manuscript age was centered around the mean chronological age of the year-group. Although age itself was a significant predictor of math ability ($\beta = 0.25$, $p < .001$), the *Numicon* strategy (using or not using) was not ($\beta = .02$, $p = .28$; see Table 4). We ran a Bayes Factor with the same prior as in the ms and yielded a Bayes Factor of $BF = 0.15$ allowing us to accept the null hypothesis.

Table 4

Numicon as a predictor of mathematics ability. Chronological age is mean centered, within each school year.

	<i>B</i>	SE <i>B</i>	β	<i>p</i>
Step 1				
Constant	0.02	0.02		.451
Age	0.73	0.06	.25	<.001
Step 2				
Constant	0.01	0.02		.451
Age	0.73	0.06	.25	<.001
<i>Numicon</i> <i>Integration</i>	0.08	0.08	.02	.280

Note: $R^2 = .06$ for step 1; $R^2 = .06$ for step 2

Numberjacks in Numerosity. Our results show that age was again a significant predictor of numerosity performance, but *Numberjacks* was not (see Table 5). As in the ms we did not have a suitable uninformed prior therefore we used an uninformed prior and found JZS Bayes Factor was 0.04 lending support for the null hypothesis.

Table 5

Numberjacks as a predictor of numerosity ability. Figures are based on 1000 bootstrap samples. Chronological was entered as years in decimals.

	<i>B</i>	SE B	β	<i>p</i>
Step 1				
Constant	72.99	1.67		.001
Age	1.57	0.19	.16	.001
Step 2				
Constant	72.94	1.67		.001
Age	1.57	0.19	.16	.001
<i>Numberjacks</i>	0.23	0.71	.006	.748
Integration				

Note: $R^2 = .02$ for step 1; $R^2 = .02$ for step 2

Numberjacks in Curriculum Math. Our results showed that age was again significant predictor for math but *Numberjacks* was not significant (see Table 6). A Bayes factor of 0.10 confirmed our support for the null hypothesis.

Table 6

Numberjacks as a predictor of mathematics ability. Chronological age is mean centered, within each school year.

	<i>b</i>	SE B	β	<i>p</i>
Step 1				
Constant	-0.001	0.02		.968
Age	0.72	0.06	.24	<.001
Step 2				
Constant	0.002	0.02		.908
Age	0.72	0.06	.24	<.001
<i>Numberjacks</i>	-0.03	0.07	-.009	.624
Integration				

Note: $R^2 = .06$ for step 1; $R^2 = .06$ for step 2

Chapter 5

Numeracy Skills in Child Synaesthetes: Evidence from grapheme-colour synaesthesia

Chapter Summary

In Chapter 4 we found that non-synaesthetes who internalised the colours of a number-colour educational tool had improved numerosity, but not improved mathematics scores. We used a dual-coding model to account for these results, and predicted that children with coloured numbers from other sources (e.g., synaesthesia) would show similar benefits. In Chapter 5 I therefore turn to synaesthesia, and ask whether there are similar differences between grapheme-colour synaesthetes and non-synaesthetic controls in their numerical cognition. Here, we test children on the same numerosity and mathematical tests and explore whether this dual-coding model applied in Chapter 4 can also apply to synaesthetes. This chapter has been prepared in paper format as Rinaldi, L.J., Smees, R, Carmichael, D. C., & Simner, J (2019) *Numeracy Skills in Child Synaesthetes: Evidence from grapheme-colour synaesthesia*. Manuscript in preparation. Note that where additional models were prepared in a supplementary information for our article submission, here they have been instead provided at the end of the chapter.

Abstract

Grapheme-colour synaesthesia is a neurological trait that causes lifelong colour associations for letter and numbers. Synaesthesia studies have demonstrated differences between synaesthetes and non-synaesthetes in ways that extend beyond synaesthesia itself (e.g., differences in their cognition, personality, and creativity). This research has focused almost exclusively on adult synaesthetes, and little is known about the profiles of synaesthetic children. By and large, findings suggest advantages for synaesthetes (e.g., Chun & Hupé, 2016; Havlik et al., 2015; Rothen et al., 2012; Rouw & Scholte, 2016; Simner & Bain, 2018) although differences in mathematical ability are unclear: some research indicates advantages (e.g., Green & Goswami, 2008) whilst others suggest difficulties (e.g., Rich et al., 2005). In the current study, we tested numerical cognition in a large group of children with grapheme-colour synaesthesia. Synaesthetes with coloured numbers showed advantages over their peers in their sense of numerosity, but not in their curriculum mathematics ability. We discuss our findings in the context of models for synaesthesia, and relate our findings, also, to wider educational practices of using coloured number-tools in schools (e.g., *Numicon*; Oxford University Press, 2018).

Introduction

Synaesthesia is an unusual neurological trait affecting at least 4.4% of the population (Simner et al., 2006). People with synaesthesia experience common stimuli (e.g., words, music) as triggering secondary experiences like colours or tastes (for review, see Simner & Hubbard, 2013). In the current study we focus on a prevalent type of synaesthesia triggered by reading. For *grapheme-colour synaesthetes*, letters and numbers give rise to automatic colour sensations. For example, a grapheme-colour synaesthete might feel that *F* is blue, *6* is red, and so on (Simner, Glover, et al., 2006; Simner & Holenstein, 2007). Grapheme-colour synaesthetes have recognised neurological differences; for example, differences in white matter connectivity in regions associated with colour processing (e.g., Rouw & Scholte, 2007) as well as differences in more distributed areas such as the superior parietal cortex (for review, see Rouw, Scholte, & Colizoli, 2011). In our study we ask whether children with grapheme-colour synaesthesia show behavioural differences to their peers in ways that extend beyond the synaesthetic sensations themselves. We compared randomly sampled child synaesthetes aged 6-10 years, and matched controls, in terms of their abilities in numerical cognition.

In adults at least, there is mounting evidence to suggest that synaesthetes have a particular cognitive profile in which they outperform non-synaesthetes in a number of ways. For example, adult grapheme-colour synaesthetes show better memory than non-synaesthetes for word lists (Gibson, Radvansky, Johnson, & McNerney, 2012), better memory for colour (Yaro & Ward, 2007), and more vivid visual mental imagery in self-report (Barnett & Newell, 2008). Rouw and Scholte (2016) found, too, that a group of synaesthetes also outperformed non-synaesthetes in a general intelligence test (and many of these synaesthetes had coloured graphemes). Similarly Chun and Hupé (2016) found that a similar group of synaesthetes were significantly better than controls in a verbal comprehension task. Ward, Thompson-Lake, Ely and Kaminski (2008) also showed that synaesthetes outperformed non-synaesthetes in objective measures of creativity, such as a convergent creativity task (i.e., finding the missing link between three ostensibly unrelated words), and that synaesthetes engage more than controls in creative activities and employment (see also Rich et al., 2005; Rothen & Meier, 2010b). In summary, synaesthetes perform better than their peers in a number of measures, suggesting they have particular differences in domains outside synaesthesia itself.

In relatively recent work, the cognitive differences found in adult synaesthetes are also now being researched in children. Simner and Bain (2018) found that child grapheme-colour synaesthetes aged 10-11 years showed superiority in a task that required them to quickly discriminate between different objects in an array. Simner and Bain (2018) also re-analysed data from a sample of grapheme-colour synaesthetes studied by Green and Goswami (2008) which pointed towards a possible verbal comprehension benefit for child synaesthetes (see also Smees et al., 2019). Together, these findings suggest that cognitive abilities may be potentially superior in synaesthetes from a young age. And certainly by the time they are adults, synaesthetes show a range of cognitive advantages over their peers.

In the current study we investigated whether differences in the cognitive profile of synaesthetes extends to numeracy skills, and in particular whether differences are found in synaesthetic children. In previous research, results on numerical cognition have been somewhat conflicting – in both adults and children. Studies have suggested that synaesthetes may experience both advantages *and* disadvantages in mathematics, depending on the type of synaesthetes tested and the way numeracy was explored. Green and Goswami (2008) measured numeracy skills in children using the WISC arithmetic test (O'Donnell, 2009), and found that grapheme-colour synaesthetes were trending towards superior scores (see Simner and Bain, 2018, for a statistical analysis of the descriptive data presented by Green and Goswami). However, the child synaesthetes tested by Green and Goswami had not been randomly sampled, and Simner and Bain (2018) describe ways in which these sampling methods may have encouraged superior performers, irrespective of whether children had synaesthesia or not. Other studies have suggested that grapheme-colour synaesthesia might in fact hinder numerical cognition. Rich, Bradshaw, and Mattingly (2005) simply asked adult synaesthetes (the majority of whom experienced grapheme-colour synaesthesia) about their experiences of mathematics; 4.7% felt they had advantages in mathematics, while 16% felt they experienced difficulties. However, there were no baselines against which to compare these responses (e.g., no groups of non-synaesthetes). And in children, Green and Goswami (2008) tested whether child grapheme-colour synaesthetes aged 7-15 would experience difficulties if numbers were presented to them in incongruent colours (i.e. colours conflicting with each child's synaesthesia) compared to congruent colours (i.e. colours matching each child's synaesthesia). Synaesthetes performed a simple digit-recall

task and showed worse memory for incongruent trials compared to a neutral baseline (black text; but see Simner & Bain, 2018, who did not replicate this finding). A similar study by Mills, Metzger, Foster, Valentine-Gresk and Ricketts (2009) looked at a case study of adult grapheme-colour synaesthesia and showed that arithmetic, too, was slower when digits were presented in incongruent colours. These latter studies suggest that grapheme-colour synaesthetes may experience difficulties from conflicting colours, but does not speak to number cognition more generally.

Finally, Ward, Sagiv, and Butterworth (2009) looked at numeracy in another type of synaesthesia altogether. *Sequence-space synaesthetes* experience sequences such as numbers as being arranged in specific spatial patterns (e.g., they may feel that numbers unfold in lines across the visual field, or wrap around the body). Sequence-space synaesthetes were slower in mental calculation for functions such as multiplication, suggesting they might ‘over-rely’ on their visuo-spatial mental number line for numerical tasks that usually involve verbal recall (e.g., multiplication; see Dehaene & Cohen, 1995, 1997; Lee & Kang, 2002). In summary, this body of research suggests that some adult grapheme-colour synaesthetes self-report difficulties in maths, that adult sequence-space synaesthetes are slower in some domains of arithmetic, and that children and adults may struggle if using coloured numbers that clash with their synaesthesia. At the same time, the adult self-report was somewhat mixed, with reports of both advantages and disadvantages in maths, and no baseline to compare against. Additionally, the child finding did not replicate using improved recruitment methods (see Simner & Bain, 2018), so it remains unclear exactly whether and how children with synaesthesia might show differences in their numerical skills. We therefore explore this topic in the current study.

Here, we tested children with synaesthesia, and measured numeracy in two ways: using a curriculum mathematics test, and a numerosity task. Numerosity is our intuitive “number sense” which allows us to understand magnitudes without counting the exact amount. Our sense of numerosity relies on an *approximate number system* (ANS) which comprises a set of mental processes that approximately encode magnitudes (Dehaene, 2001). Numerosity is often measured by asking individuals to make quantity judgements without enough time to physically count objects (Dehaene, 2001). For example, in the *Dot Numerosity* task used here (Halberda Mozzocco & Feigenson, 2008), children view a cluster of black dots adjacent to a cluster of white dots. Both appear on the screen simultaneously for a short period of time. Children must then decide which array

contained more dots. Adults are typically able to perform this task successfully, and can differentiate between dot arrays with a ratio of 1:1.15 (Barth et al., 2003). In children, Lipton and Spelke (2004) found that even 6-months-olds discriminate at a ratio of 1:2, and that by 9 months, infants had improved this to somewhere between 1:1.5 and 1:1.25. The ANS therefore develops over time but is already established in young babies (see also Feigenson, Dehaene, & Spelke, 2004; Xu & Spelke, 2000).

In our study we look at how synaesthetes perform in a test of numerosity, as well as a traditional curriculum maths test. Evidence suggests there is some interaction between both types of numerical cognition, since higher numerosity performance is linked with higher maths scores (Anobile, Stievano, & Burr, 2013; Chen & Li, 2014; Halberda et al., 2008). For example, Halburda et al. (2008) investigated children's numerosity ability and maths performance, and showed that differences in numerosity at age 14 is correlated to mathematics performance as far back as kindergarten. Wong, Ho, and Tang (2016) used structural equation modelling to suggest a causal directionality, in that better numerosity leads to improved numeral mapping, which consequently leads to improved mathematical skills. In the present analyses we investigate both numerosity and maths performance, asking first whether synaesthetes have superior performance in numerosity (ANS acuity), and then whether they also show superior performance in maths. We predict that child synaesthetes may perform differently to controls in tests of numerosity and/or mathematics compared to non-synaesthetic controls, and we review the basis of this hypothesis below.

One recent study has suggested that pairing colour with numbers could be tied to advantages in numerosity, even for non-synaesthetes. Rinaldi, Smees, Alvarez, and Simner (2019) looked at the pairing of colour with number in educational maths tools such as *Numicon* (Oxford University Press, 2018). *Numicon* consists of ten colour-coded plastic shapes, corresponding to the numbers 1-10 (e.g., the shape for number 5 has five holes and is coloured red). Rinaldi et al. looked at how colour-coding in this tool aided children's learning. They tested a large cohort of children who had been taught with *Numicon* at school, and divided children into two groups: those who had naturally memorised the colour-coding of *Numicon*, versus those who had not. Rinaldi et al. found that children who had internalised *Numicon* colours (e.g., 5 is red) performed better in a dot numerosity task compared to their peers who had not internalised these colours. Rinaldi et al. suggested that the 'dual coding' of colours to numbers may have

strengthened children's numerical encoding, leading to a stronger ANS and therefore improved numerosity skills. This type of dual-coding model was originally proposed by Paivio (1969), but has since been offered within models of synaesthesia (e.g., Gibson et al., 2012). Children who encoded *Numicon* colours in this earlier study were not synaesthetes, but they show synaesthesia-like associations¹² suggesting that genuine grapheme-colour synaesthetes, too, might benefit from coloured numbers in a similar way.

If applied to synaesthetes, the *Numicon* findings of Rinaldi et al. would predict that children with grapheme-colour synaesthesia might show benefits in numerosity, but no benefits in mathematics. This is because children who internalised *Numicon* colours were better than controls in numerosity, but not in a curriculum maths test. We therefore present both numerical tasks to our grapheme-colour synaesthetes. Here we will compare synaesthetes with colours only for letters, to synaesthetes with colours for numbers. Splitting our grapheme-colour synaesthetes in this way allows us to test the dual-coding model of Rinaldi et al. (2019): dual-coding predicts numerical benefits for synaesthetes with coloured numbers but not coloured letters. However, if synaesthetes score well for reasons beyond dual-coding (e.g., from some broader type of enhanced perceptual or structural organisation (see Hänggi, Wotruba, & Jäncke, 2011; Ramachandran & Azoulay, 2006; Simner & Bain, 2018), we might find higher numerosity (and potentially even, better curriculum maths scores) irrespective of whether synaesthetes have colours from letters or from numbers.

We also tested two types of non-synaesthetes as controls: non-synaesthetes with *average-memory* for multisensory stimuli, and non-synaesthetes with *superior-memory* for multisensory stimuli (see *Methods*). These latter can recall coloured-graphemes in short-term memory tests particularly well (i.e., they can invent colours for numbers/ letters, then recall these associations a few minutes later), but they do not have the life-long

¹² Rinaldi et al. tested children from the same population as our study here, but their target population (children internalising *Numicon*) were very different to our targets here (children with synaesthesia). Synaesthetes have largely idiosyncratic colours, while *Numicon-internalizers* have a fixed set of colours, following the maths tool. And synaesthetes are identified very differently: they must consistently report their colours in retests across periods as long as approximately 7 months (see *Methods*), while *Numicon-internalisers* simply state their colours once (and match to *Numicon* at rates higher than chance). When comparing children from both target groups (i.e., current study vs. Rinaldi et al., 2019), only two in 41 synaesthetes we test here appeared in both groups. These two children cannot be ruled out as legitimate synaesthetes, since children with synaesthesia can, on rare occasions, “imprint” their colours from the environment (Witthoft & Winawer, 2006). However, for clarity we point out that removing these two children from our current study does not alter the pattern of results reported below in any way.

associations found in synaesthetes. Including both *high-* and *average-memory controls* in our study allows us to unpack whether benefits for synaesthetes in numerical cognition relate in any way to having a good memory – in which case *high-memory controls* might perform as well as synaesthetes. Conversely, if synaesthetes have advantages unrelated to this type of memory ability, they should out-perform both groups of controls.

Methods

Participants

We tested 34 children with grapheme-colour synaesthesia who had been identified from an earlier screening program (Rinaldi, Smees, Carmichael, & Simner, 2019b; Simner, Alvarez, Rinaldi, et al., 2019; Simner, Rinaldi, et al., 2019). This program identified child synaesthetes between the ages of 6 and 10 years, from the student bodies of 22 UK primary schools in the south of England, Years 2 through 5. Since opt-outs were minimal (approximately 1%), this sample represents an unbiased cohort of local child synaesthetes. The screening methodology is described fully within Rinaldi et al. (2019; see also; Simner, Alvarez, Rinaldi, et al., 2019; Simner, Rinaldi, et al., 2019) but essentially required each child to repeatedly pick colours for the letters A-Z and numbers 0-9 from an extensive colour-palette. Synaesthetes were identified by detecting the gold standard characteristic of ‘consistency over time’ (i.e., for a genuine synaesthete, associations tend to stay the same over time; e.g., if the letter *A* is red, it is *always* red). To be identified as a synaesthete, a child therefore had to be statistically more consistent than age-matched peers when reporting his/her grapheme-colour associations in three comparisons: within an initial consistency test (Session 1), *and* within a second consistency test (Session 2), *and* across the 7 months between these two sessions. In other words, their methods for identifying synaesthetes were highly conservative, and full details are given in Rinaldi et al. (2019; see also; Simner, Alvarez, Rinaldi, et al., 2019; Simner, Rinaldi, et al., 2019). Once synaesthetes were identified, we divided them into two groups: synaesthetes with only coloured letters ($n = 14$), versus synaesthetes who had coloured numbers ($n = 20$, including 13 synaesthetes who had both letters and numbers). Henceforth we refer to

these as *letter-only synaesthetes*, and *number-synaesthetes* respectively¹³. We identified and excluded an additional 7 children who have been identified with grapheme-colour synaesthesia but also had yet another type of synaesthesia (which triggered sensations *other* than colour). Since we were interested in colour specifically, we did not include these children within our study. Full demographic details of our final groups are given in Table 1.

In addition to synaesthetes, we also tested non-synaesthetic controls. These controls were children drawn from the same population as synaesthetes, but had failed the synaesthesia diagnostic. We divided our controls into two groups: both were non-synaesthetes but they differed in one element of the screening test. *Average-memory controls* performed within the average range within the Session 1 consistency test, whereas *high-memory controls* were superior performers in Session 1 (although they did not maintain consistency in Session 2 or across Sessions). High-memory controls therefore showed an increased ability to remember paired associations (e.g., colours for numbers within a single test session) but without having the long-term consistency characteristic of synaesthesia. Comparing both types of controls with synaesthetes will therefore allow us to distinguish features of synaesthesia from considerations of memory (see Simner & Bain, 2018; Simner et al., 2009).

Average-memory controls were matched pairwise to each synaesthete and to each *high-memory control* (in an approximate ratio of 2:1) in both age and sex, and also, where possible, within schools. Where school-matching was not possible, controls were matched from a school sharing the same socio-economic status (i.e., using each school's percentage *Free School Meals*, as the UK school-wide benefit linked to low household income; see Taylor, 2018). All children (Years 2-5) completed our numerosity test, while only Years 3-5 completed our maths test (see Methods for details). An additional 15 participants were tested but subsequently excluded from our numerosity analysis: nine children experienced a technical failure and six children did not finish the task.

¹³ Our crucial focus is whether synaesthetes have coloured numbers or not. Due to limited numbers of synaesthetes, we collapsed two group of synaesthetes together: those with coloured numbers only, and those with coloured numbers and letters. These children all had coloured numbers, so formed the 'number-synaesthetes' group. Our comparison group of synaesthetes had *no* coloured numbers (i.e., 'letter-only synaesthetes').

Table 1

Number (N) of participants by group and gender, mean age and standard deviation (SD). For each analysis, we compare each type of *synaesthete* to *high- memory controls* and their respective *average-memory controls* (e.g., the analysis for *letter-only synaesthetes* will compare participants in row 2, row 4, and combined rows 6 and 8).

Group	Total	N	N	Mean	SD
	N	Female	Male	age	age
Letter-only synaesthetes	14	9	5	8.64	1.30
Number synaesthetes	20	10	10	8.88	1.15
High-memory control	242	125	116	8.27	1.22
Average-memory control:					
- matched to letter-only synaesthetes	28	18	10	8.60	1.23
- matched to number synaesthetes	40	20	20	8.85	1.18
- matched to high-memory controls	513	268	245	8.31	1.19

Materials and Procedure

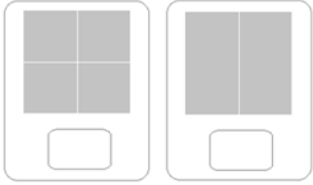
Our study received ethical approval by the Sussex University Science and Technology Research Committee. Children completed two tests of numerical cognition, described below. Testing took place between October 2016 and April 2017. Neither children nor experimenters knew the synaesthetic status of children at the point of their testing for numerosity and mathematics.

Numerical Cognition: Curriculum Maths.

Our in-house maths test came from Rinaldi, Smees, Alvarez et al. (2019), and assessed key components from the UK primary school mathematics curriculum (“The national curriculum in England: Key stages 1 and 2 framework document,” 2013). Our pencil-and-paper test had 47 questions in total, which represented one question for each of the 7-9 topics per year – across six school years (Years 1-6). These topics covered a range of subjects including arithmetic, fractions, percentages, geometry and so on (see Figure 1 for examples from the test). Children started the test with questions two years below their current school year (e.g., Year 3 students start with Year 1 questions). Since there is no set UK math curriculum prior to Year 1, students in Year 2 could not complete an equivalent test so were excluded from mathematics testing. The test presented one

question per line, and children were given five minutes to answer as many questions as possible. Children were not expected to go beyond their current year group material in the allocated time, although all correct questions were scored.

1- Tick the picture that shows quarters



4- Please subtract

$$\begin{array}{r} 9,547 \\ - 4,369 \\ \hline \end{array}$$

2- Fill in the missing number below

30

40

50

60

70

90

5- Calculate the area of the rectangle below.

2 cm

4 cm

cm²

3- Tick the box where 8 has a value of eighty?

813 ☐

278 ☐

84 ☐

6- Please calculate

15% of 420 =

Figure 1. Example questions from each year of the curriculum maths test (curriculum year shown in grey).

Numerical Cognition: Dot Numerosity.

Our numerosity task was presented on electronic tablets. Children were each given a touch screen Acer Aspire SW3-016 or Acer One 10 tablet, which ran on Intel® Atom TM x5-Z8300 Processors, with Windows 10 and had 10.1" LED backlight touchscreens (1280 x 800 pixels). As in Rinaldi, Smees, Alvarez et al. (2019), our task was the Panamath dot numerosity task (Halberda et al., 2008), which we presented with a task-time of 2 minutes, and default settings which generate an adjusted level of difficulty based on each child's age (entered in whole years). This test briefly presents a cluster of white dots adjacent to a cluster of black dots (1382 – 1951ms dependent on age; see Figure 2 for screen-shot). Children were required to press one of two buttons (marked with a white or black sticker) to indicate whether there had been more white dots or black dots. Children were told they would play a short game in which they would not have time to count the dots, but should make their best guess as quickly as possible.

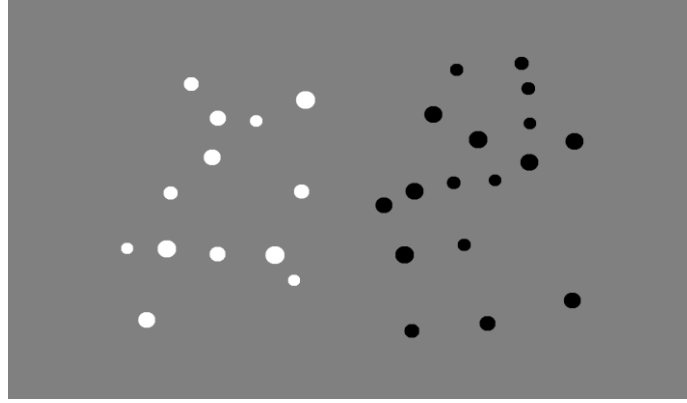


Figure 2. A screenshot of the Panamath dot numerosity test.

Results

We examined differences in numeracy skills, and first compare *letter-only synaesthetes* to controls (i.e., we compare *letter-only synaesthetes*, *high-memory controls*, and *their average-memory controls*), then repeat the process for *number synaesthetes*. We include age as a covariate in our models given that synaesthetes and *high-memory controls* were not age-matched to each other, and there is a known limitation in this regard for 6-year old synaesthetes¹⁴. Where appropriate we present mixed effects models, which are widely used with nested data (see Field, Miles, & Field, 2012) to capture random effects caused by different classes within different schools. We ran our Linear Mixed Effects models in R version 3.5.0 (R Core Team, 2016) using *lme4* (Bates, Maechler, Bolker, & Walker, 2015) and using *lmerTest* to obtain *p*-values (Kuznetsova, Brockhoff, & Christensen, 2017). Unless otherwise stated, we set our largest cohort as the reference group (i.e. *average-memory controls*, but see *Supplementary Information (SI)* at the end of the chapter for parallel models switching reference group to *high-memory controls*).

¹⁴ The diagnostic used in this study to identify synaesthetes has a known limitation for 6-year olds. Six year old synaesthetes have only very nascent synaesthesia (Simner et al., 2009) and the diagnostic can detect only those synaesthetes with most synaesthetic colours (typically the older of the 6 year olds). This weights 6 year olds away from being diagnosed as synaesthetes, and towards being diagnosed as high-memory non-synaesthetes (see Simner, Alvarez, Rinaldi, et al., 2019; Simner, Alvarez, Smees, et al., 2019). No age effects are found at other ages, where the test performs better. Given this age effect in 6 year olds, we included age in our model as a co-variate. Finally, we point out that this age-influence in our diagnostic makes our comparisons here more conservative (i.e., some 6 year old synaesthetes are pushed into the high memory group, making group-wise differences harder, not easier, to detect).

Do synaesthetes show differences in numerosity?

Following Rinaldi, Smees, Alvarez et al. (2019), we analysed percent correct responses on the numerosity task. Scores notably lower than chance ($<45\%$) were removed because this suggested confusion with key-bindings. We therefore removed 3 *high-memory controls* and 3 *average-memory controls* in our letter-only analysis (leaving 14 *synaesthetes*, 239 *high-memory controls* and 538 *average-memory controls*). In our analysis for *number synaesthetes* we removed 3 *high-memory controls* and 4 *average-memory controls* (leaving 20 *synaesthetes*, 239 *high-memory controls* and 549 *average-memory controls*). Our percent correct variable was skewed with most children performing well on our task. We consequently used bootstrapped models. In an initial test we found no random effects of class or school (i.e., Linear Mixed Effects analysis not required), so we report bootstrapped linear regression models with covariates of age (i.e., age at test, in years and decimals) and gender.

Letter-only Synaesthetes

We first looked at whether there were any significant differences between *letter-only synaesthetes* and controls. We found a significant age effect, but no significant difference between *letter-only synaesthetes* and our *average-memory controls* in correct number of responses, and no significant differences between *high-memory* and *average-memory controls* (Table 2). We switched our reference to *high-memory controls* and found the same; no differences between *letter-only synaesthetes* and *high-memory controls* (Table 1 in SI).

To explore our null result, we produced a Bayes Factor to determine whether we have enough evidence to accept the null hypothesis (Dienes, 2014). We used an uninformative prior using the *BayesFactor* package in R (Morey, Rouder, & Jamil, 2015). Bayes Factors lie on a continuum, where scores of less than 0.33 provide evidence for the null hypothesis and scores above 3 provide evidence for the experimental hypothesis (Dienes, 2014). Here we found a JZS Bayes Factor of 0.31 suggesting we have sufficient evidence to accept the null hypothesis. This means that there is no difference between letter-only synaesthetes and controls in numerosity.

Table 2

Group status (*letter-only-synaesthetes, high-memory controls*) as a predictor of numerosity with *average-memory controls* as reference and based on 1000 bootstrapped samples. Chronological age is age in decimals.

	Estimate(B)	SE (B)	<i>p</i>	<i>95% CI</i>	
Step One					
Constant	77.59	2.63	.001	72.38	82.87
Age	1.18	0.30	.002	0.58	1.78
Step Two					
Constant	77.42	2.65	.001	71.89	82.63
Age	1.18	0.30	.002	0.59	1.76
Letter-only-synaesthetes (vs. average-memory controls)	1.12	2.14	.596	-3.13	5.33
High-memory Controls (vs. average-memory controls)	0.54	0.76	.503	-0.96	2.04

Note: $R^2 = .019$ for step 1; $R^2 = .020$ for step 2

Number Synaesthetes

We repeated our analysis with *number synaesthetes*. Here, we found that synaesthetes significantly out-performed *average-memory controls* ($p = .001$; see Table 3), and *high-memory controls* ($p = .015$; see Table 2 SI) with *number synaesthetes* on average scoring 3.8% higher in percent correct numerosity scores than *average-memory controls* and 3.5% higher than *high-memory controls*; this data is shown in Figure 5.

Table 3

Group status (*number-synaesthetes, high-memory controls*) as a predictor of numerosity with *average-memory controls* as reference and based on 1000 bootstrapped samples. Chronological age is age in decimals.

	Estimate(B)	SE (B)	<i>p</i>	<i>95% CI</i>	
Step One					
Constant	77.35	2.51	.001	72.58	82.26
Age	1.23	0.28	.001	0.68	1.76
Step Two					

	Estimate(B)	SE (B)	<i>p</i>	95% <i>CI</i>	
Constant	77.36	2.50	.001	72.63	82.31
Age	1.20	0.28	.001	0.67	1.73
Number-synaesthetes	3.24	1.01	.001	1.31	5.33
(vs average-memory controls)					
High-memory Controls	0.46	0.74	.538	-0.98	1.92
(vs average-memory controls)					

Note: $R^2 = .022$ for step 1; $R^2 = .024$ for step 2

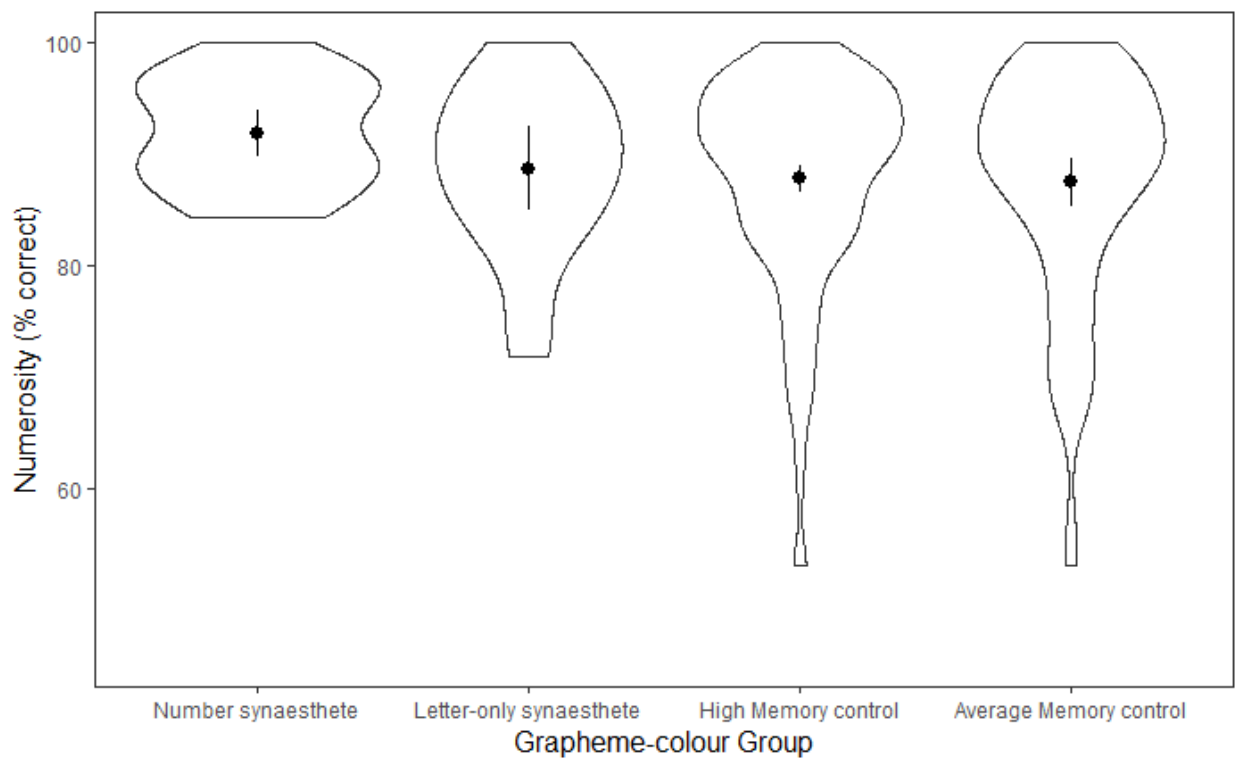


Figure 4. Violin plots illustrating the difference between grapheme-colour synaesthesia subtypes in correct numerosity responses. Error bars show 95% confidence intervals. For the purposes of illustration, all *average-memory controls* (matched to *number synaesthetes*, *letter-only synaesthetes*, and *high-memory synaesthetes*) have been combined in this figure.

Do Synaesthetes show differences in Curriculum Mathematics?

Only Years 3 to 5 took our maths test, so we examine differences in our letter-only groups between 10 *letter-only synaesthetes*, 172 *high-memory controls*, and 380 *average-memory controls*. In our *number synaesthete* groups we examine differences between 17

number synaesthetes, 172 *high-memory controls*, and 392 *average-memory controls*. Since different year groups saw different versions of the maths test (i.e. Year 3 started with Year 1 questions whereas Year 4 started with Year 2 questions) we first converted raw maths scores into z-scores standardized within year group. As with numerosity, we include age as a covariate, treating this as age-centred within year groups (since our test was based on school year rather than chronological age). Finally, we tested for, and found, random effects of class and school and therefore report a linear mixed effects (LME) model including random intercepts for class and school (see Table 3).

Letter-only Synaesthetes

Taking *average-memory controls* as the reference group, we found no difference between *letter-only synaesthetes* and *average-memory controls* in maths score ($p = .894$) but we did find that *high-memory controls* were trending higher than *average-memory controls* ($p = .069$; See Table 4). There was also a significant effect of age. We found a similar pattern when switching the reference group to *high-memory controls*; there was no difference between *letter-only synaesthetes* and *high-memory controls* (See Table 3 in SI). We again produced a Bayes Factor to confirm our null result of letter-only synaesthesia status. We found a JZS Bayes Factor of 0.19, again suggesting moderate evidence to accept the null hypothesis.

Table 4

Group status (*letter-only synaesthetes*, *high-memory controls*) as a predictor of maths controlling for random effects of school and class with *average-memory controls* as reference. Chronological age is age within year group.

Fixed Effects	Estimate(B)	SE (B)	t	$p(t)$
Intercept	-0.03	0.08	-0.42	.678
High-memory control	0.15	0.08	1.82	.069
(vs average-memory control)				
Letter-only synaesthetes	-0.04	0.30	-0.13	.894
(vs average-memory control)				
Age	0.51	0.14	3.73	<.001***
Random Effects	Variance	SD	X^2	$p(X^2)$
Class	0.11	0.34	29.10	<.001***
School	0.04	0.21	12.72	<.001***
Residual	0.79	0.89	-	-

Number Synaesthetes

Again taking *average-memory controls* as the reference, we found no difference between *synaesthetes* and *average-memory controls* in maths scores ($p = .629$), but *high-memory controls* were trending higher than *average-memory controls* ($p = .070$; See Table 5), and we found a significant age effect. Again when we switched the reference to *high-memory controls* we found a similar pattern: no difference between *synaesthetes* and *high-memory controls* (See Table 4 in SI). We again produced a Bayes Factor to confirm our null result in *number synaesthetes*, again using the *BayesFactor* package in R. We found a JZS Bayes Factor of 0.16 again suggesting moderate evidence to accept the null hypothesis.

Table 5

Group status (*number synaesthetes*, *high-memory controls*) as a predictor of maths controlling for random effects of school and class with *average-memory controls* as reference. Chronological age is age within year group.

Fixed Effects	Estimate(B)	SE (B)	t	$p(t)$
Intercept	-0.04	0.08	-0.55	.589
High-memory control	0.15	0.08	1.82	.070
(vs average-memory control)				
Number synaesthetes	0.11	0.23	0.48	.629
(vs average-memory control)				
Age	0.44	0.13	3.31	<.001**
Random Effects	Variance	SD	X^2	$p(X^2)$
Class	0.12	0.34	34.86	<.001***
School	0.05	0.23	14.06	<.001***
Residual	0.79	0.89	-	-

Discussion

Here we examined the numerical cognition of children 6-10 years with grapheme-colour synaesthesia (i.e., lifelong associations of coloured letters or numbers). We compared two

types of grapheme-colour synaesthetes to two types of non-synaesthetic controls. We compared synaesthetes with and without number associations (i.e. *number synaesthetes* and *letter-only synaesthetes*) to both *high-memory controls* (i.e., children who can recall similar associations very well in the short-term, but are not synaesthetes) and *average-memory controls* (i.e., children with average recall in this domain). We found that *number synaesthetes* performed significantly better than both types of controls in a numerosity task (i.e., estimating which of two dot-clusters was more numerous, with only brief exposure). However, we found no differences between *letter-only synaesthetes* and controls in numerosity, and we found no difference between either type of synaesthete and controls in mathematics.

Importantly, we highlight here that the numerosity advantage was present only in synaesthetes with coloured numbers, suggesting support for a dual-coding account (Paivio, 1969). In this type of model, colour-coding numbers could provide more robust representations, and therefore strengthen numerical cognition. It is important to note that this finding was limited to *synaesthetes* (with their lifelong colour associations) but was not found for *high-memory controls* (who can easily generate and remember similar associations, but only in the short-term). This suggests that improvements in numerosity come from colour associations that are robust and long-term.

A similar finding has emerged from a group of children who internalised colours for numbers, but were not synaesthetes (Rinaldi, Smees, Alvarez, et al., 2019). As noted in our Introduction, these children learned coloured numbers from the educational tool *Numicon*, and showed advantages in the same test of numerosity over their peers who had not memorised colours (even though all children had been exposed to *Numicon*). Across both studies, we might therefore infer that dual-coding of numbers improves numerosity – whether synaesthetic or not. But how does this advantage in numerosity come about? One answer may lie in how colours encode into the mental number system. Rinaldi et al. concluded that numerosity advantages require colours to be associated to magnitude, and not simply to Arabic numerals. They drew this conclusion by tracing their numerosity finding to number tools that colour-code magnitude in particular – such as *Numicon* (which pairs colours to plastic shapes with holes denoting magnitude; a comparison tool linking colours to the shapes of Arabic numerals did not show a similar effect). Importantly, there is some evidence that colours target magnitude in synaesthesia, too (Berteletti, Hubbard, & Zorzi, 2010; Gertner, Arend, & Henik, 2013; Kadosh et al., 2005).

For example, Berteletti et al. (2010) presented a ‘synaesthetic Stroop’ task in which a synaesthete case-study saw coloured numerals or dot patterns (this latter representing magnitude). The synaesthete had to ignore the number but name the ink colour, and showed a congruency effect (faster responses when colour matched synaesthesia). Importantly, this was true for both digits *and* dot-patterns, suggesting that synaesthetic colours attach to magnitudes (i.e., not just to numerals) and may thereby influence numerosity judgements. In combination with Rinaldi, Smees, Alvarez et al. (2019), we therefore have evidence across two different groups (*synaesthetes* and *non-synaesthetes*) that internalising colours for numbers can associate with superior scores in numerosity – especially if those colours are encoded at the level of magnitude.

Overall our findings suggest that child synaesthetes show improved numerosity skills, but this did not translate into improved mathematics skills. Numerosity has a well-documented relationship with maths (Halberda et al., 2008; Wong et al., 2016) but synaesthetes benefiting at one level did not benefit at the other. This was true not only in our own data, but also in Rinaldi et al. (2019). This suggests that whatever colour-benefits are enjoyed by the Approximate Number System in numerosity, do not propagate through to processes governing mathematics. The reasons for this are unclear. It may be that improvements in numerosity were simply not strong enough resonate through to mathematics. Alternatively it may simply be that our in-house mathematics test was not sensitive enough to detect them – an important possible limitation of our study.

We note here that there have been concerns about the relationship between numerosity dot tasks and inhibitory control. Inhibitory control is the ability to suppress salient but task-irrelevant information (Merkley, Thompson, & Scerif, 2016). A common example is the Stroop task: you must be able to ignore the salient yet task-irrelevant meaning of colour-words when naming their font-colour (Stroop, 1935). Concerns regarding dot-numerosity tasks pertain to how dots appear on-screen. In order to compensate for the fact that fewer dots would take up less surface area, the size of the dots is manipulated (i.e., where dots are fewer, they are also larger). This conflicting information (ignore dot-size, estimate only their number) introduces inhibitory control requirements (Clayton & Gilmore, 2015). We might therefore question whether synaesthetes are better at numerosity, or better at inhibitory control. Here we tentatively suggest the former, for several reasons. First, it would be unclear why *number-synaesthetes* and *letter-only synaesthetes* should differ in inhibitory control abilities (i.e., a numerosity account is

more logical). Second, recent studies (Malone et al., 2019) have questioned the role of inhibitory control in numerosity tasks, since they find no correlation between independent measures of inhibitory control, and maths ability.

In summary, we have shown that children with grapheme-colour synaesthesia have superior numerosity scores, tied to coloured numbers in particular. The nature of regression statistics do not allow us to infer the direction of causality but we have tacitly assumed that synaesthetic colours for numbers improve sense of magnitude. However, we acknowledge that the reverse might also be true: children with better numerosity may be better able to integrate colours into their magnitude schema and thereby develop synaesthesia. Our evidence supports a dual-coding account and joins a literature where synaesthetes benefit via a range of mechanisms (both dual-coding and otherwise). For example, although grapheme-colour synaesthetes have superior numerosity if their numbers are coloured (but not letters), they also show broader advantages for stimuli such as faces or scenes (Gross, Nearing, Caldwell-Harris, & Cronin-Golomb, 2011; Pritchard, Rothen, Coolbear, & Ward, 2013; Rothen & Meier, 2010a; Ward, Hovard, Jones, & Rothen, 2013). Broad advantages do not negate the possibility of dual-coding, because more than one mechanism may work in parallel. These parallel mechanisms might perhaps be differences in “cognitive processing style” (Meier & Rothen, 2013a), or “enhanced perceptual organisation” (Hänggi et al., 2011; Ramachandran & Azoulay, 2006; Simner & Bain, 2018) although neither have been fully elaborated. In conclusion, our data join findings elsewhere in the literature, showing the range of cognitive benefits enjoyed by synaesthetes – which we can now extend to benefits in numerosity tasks.

Chapter 5: Supplementary Information

Additional models switching the reference group to high-memory controls

Do Synaesthetes show differences in Numerosity?

Letter-only Synaesthetes

We investigated differences between *letter-only synaesthetes* and *high-memory controls*. We found a significant effect of age, but no significant difference between synaesthetes and *high-memory controls*, and no significant differences between *average-memory controls* and *high-memory controls* (See Table 1)

Table 1

Group status (*letter-only synaesthetes*, *average-memory controls*) as a predictor of dot numerosity with *high-memory controls* as reference and based on 1000 bootstrapped samples. Chronological age is age in decimals.

	Estimate(B)	SE (B)	<i>p</i>	<i>95% CI</i>	
Step One					
Constant	77.59	2.49	.001	72.48	82.61
Age	1.18	0.28	.001	0.63	1.75
Step Two					
Constant	77.96	2.52	.001	73.16	83.12
Age	1.18	0.28	.001	0.63	1.74
Letter-only synaesthetes (vs High-memory controls)	0.58	2.24	.783	-4.19	4.80
Average-memory controls (vs High-memory controls)	-0.54	0.78	.492	-2.21	0.92

Note: $R^2 = .02$ for step 1; $R^2 = .02$ for step 2

Number Synaesthetes

We ran our analyses again against *high-memory controls*. There was a significant difference age effect, and a significant difference between *number-synaesthetes* and *high-memory controls* ($p = .015$; see Table 2). There was no difference between *average-memory* and *high-memory controls*.

Table 2

Group status (*number-synaesthetes*, *average-memory controls*) as a predictor of dot numerosity with *high-memory controls* as reference and based on 1000 bootstrapped samples. Chronological age is age in decimals.

	Estimate(B)	SE (B)	<i>p</i>	<i>95% CI</i>	
Step One					
Constant	77.35	2.55	.001	72.40	82.60
Age	1.23	0.29	.001	0.63	1.78
Step Two					
Constant	77.81	2.59	.001	72.78	83.20
Age	1.20	0.29	.002	0.62	1.74
Number-synaesthetes (vs high-memory controls)	2.78	1.13	.015	0.56	5.23
Average-memory controls (vs high-memory controls)	-0.46	0.75	.549	-1.90	1.04

Note: $R^2 = .022$ for step 1; $R^2 = .024$ for step 2

Do Synaesthetes show differences in Curriculum Mathematics?

Letter-only Synaesthetes

Taking *high-memory controls* as the reference group we find no differences between *letter-only synaesthetes* and *high-memory controls* ($p = .521$; See Table 3).

Table 3

Group status (*number synaesthetes*, *average-memory controls*) as a predictor of maths controlling for random effects of school and class with *high-memory controls* as reference. Chronological age is age within year group.

Fixed Effects	Estimate(B)	SE (B)	<i>t</i>	<i>p(t)</i>
Intercept	0.12	0.10	1.26	.215
Average-memory control (vs high-memory control)	-0.15	0.08	-1.82	.069
Letter-only synaesthetes (vs high-memory control)	-0.19	0.30	-0.64	.521
Age	0.51	0.14	3.73	<.001***
Random Effects	Variance	SD	X^2	$p(X^2)$
Class	0.11	0.34	29.10	<.001***
School	0.04	0.21	12.72	<.001***
Residual	0.79	0.89	-	-

Number Synaesthetes

Again taking *high-memory controls* as the reference group we see no differences between *number synaesthetes* and *high-memory controls* ($p = .848$; See Table 4).

Table 4

Group status (*grapheme-colour synaesthetes, average-memory controls*) as a predictor of maths controlling for random effects of school and class with *high-memory controls* as reference. Chronological age is age within year group.

Fixed Effects	Estimate(B)	SE (B)	<i>t</i>	<i>p(t)</i>
Intercept	0.11	0.10	1.09	.284
Average-memory control (vs High-memory control)	-0.15	0.08	-1.82	.070
Number synaesthetes (vs High-memory control)	-0.04	0.23	-0.18	.848
Age	0.44	0.13	3.31	<.001**
Random Effects	Variance	SD	X^2	$p(X^2)$
Class	0.12	0.34	34.86	<.001***
School	0.05	0.23	14.06	<.001***
Residual	0.79	0.89	-	-

Chapter 6

General Discussion

In this thesis, I set out to investigate synaesthesia during childhood. Overall, this thesis asked, very simply, what child synaesthetes are *like*. In particular, this thesis focused on two specific areas. The first two experimental chapters investigated the personality profile of a child synaesthete. The second two experimental chapters examined numerical cognition, first in non-synaesthete children and then in synaesthete children as well. In this General Discussion, I will summarise the key findings from each of the preceding experimental chapters, relating their findings where relevant and making key suggestions for future research in each domain. I will finish by offering reflections on running large-scale studies.

Personality: What we have learnt?

In the first half of this thesis, we investigated the personality profile of child synaesthetes. Chapter 2 described the development and validation of three measures of personality in children. Here, we took existing Big Five measures of personality aimed at adolescents and adults, and created and/or validated versions of these tests for children and parents. Our validation showed that the (existing) BFI-44-Parent had the expected factor structure, good internal reliability, and convergent validity. Secondly, we adapted the existing BFI-44-A (originally for adolescents) to be suitable for children by adding definitions for words with an age of acquisition higher than our target sample. Using this method, we were able to make our new test, the Definitional-BFI-44-C, appropriate for children as young as eight years old. We again found the expected factor structure in the Definitional-BFI-44-C, after controlling for acquiescence bias, with adequate reliability and validity. Lastly, we created a short, 10-item pictorial test for yet-younger children, aged six to ten, based on the adult 10-item Big Five Inventory form. Here we found the expected five factor structure again after controlling for acquiescence bias. However, when we split our sample into younger (6-7 year old) and older children (8-10 year old), we were able to find the expected 5-factor structure (for *Openness*, *Conscientiousness*, *Extraversion*, *Agreeableness*, and *Neuroticism*) only in the older children, while the younger children

showed only convergence for *Agreeableness* and *Neuroticism*. This finding underlines the importance of accounting for age in the measurement of personality, and the potential difficulty of accessing self-report from younger children. For this new test, the Pictorial-BFI-10-C, we found adequate, though not optimal, reliability and validity. Overall, Chapter 2 successfully validated three personality questionnaires for children, especially in children aged 8-10.

In Chapter 3 we investigated what synaesthetes are like in terms of their personality profile, again using the Big Five personality traits of *Openness*, *Conscientiousness*, *Extraversion*, *Agreeableness* and *Neuroticism*. To do this, we made use of the personality instruments that were best validated in Chapter 2: The Definitional-BFI-44-C for child self-report and the BFI-44-Parent questionnaire for parent report. Here we investigated whether the adult personality profile of synaesthesia could be found in children. In previous studies, systematic differences in *Openness*, *Agreeableness*, *Neuroticism* and *Conscientiousness* have all been linked to synaesthesia (Banissy et al., 2013; Chun & Hupé, 2016; Rouw & Scholte, 2016). However, only higher levels of *Openness* have been linked to synaesthesia consistently across *all* previous adult studies, which frequently had methodological or recruitment confounds. We investigated personality in our child sample using multinomial regression models, which predicted the likelihood of membership in one of four groups: *OLP synaesthete*, *grapheme-colour synaesthete*, *high-memory control*, or *average-memory control*. Here we replicated adult findings in children: that synaesthesia, regardless of type, was associated with increased *Openness*.

Beyond this, our results in Chapter 3 and previous results in adult synaesthetes diverge. Previous results found adult synaesthetes had lower *Agreeableness* (Banissy et al., 2013) and higher *Neuroticism* traits (Rouw & Scholte, 2016). However, our findings in children did not show the same pattern. We instead found a role for *Extraversion*; i.e., grapheme-colour synaesthetes, but not OLP synaesthetes, had lower *Extraversion* than their peers. The *Extraversion* finding is previously unreported in any other study and requires further rigorous investigation in adult synaesthetes. We also suggested that there are shared brain regions implicated in the development of *Openness* and *Extraversion* traits, and that some of these shared regions are shared too, with regions associated with grapheme-colour synaesthesia, which may help to explain this finding. We additionally found evidence that OLP synaesthetes had increased *Conscientiousness* traits. However, the OLP screening task requires high levels of attention to detail; therefore, we suggested that this finding

may be a result of the OLP screening task, rather than OLP itself (i.e., those who are high in *Conscientiousness* may pay more attention to the task, and thereby be more likely to receive a synaesthesia diagnosis). Overall however, Chapter 3 showed that across a large, randomly selected cohort of child synaesthetes, particular personality traits are systematically associated with specific types of synaesthesia. Together, Chapters 2 and 3 present researchers with intriguing new findings about the personality of synaesthetes and the tools with which to continue this research.

Future research in Synaesthesia and Personality

One future avenue of research for testing personality in children more broadly (Chapter 2) may be to reconsider the underlying factor structure of the personality traits. In Chapter 2, we were unable to validate the Pictorial-BFI-10-C for our youngest children, finding that they only had stable factors of *Agreeableness* and *Neuroticism*. We may in the future re-examine this data considering a different framework. Younger children have less-defined boundaries between Big Five traits (Caspi et al., 2005; Soto et al., 2008). One reason for this may be because, as Soto et al. (2008) suggest, children are taught that behaviours are good or bad, which causes difficulties with differentiation between traits such as *Agreeableness* and *Conscientiousness* (as both correspond to “good”, when traits are high). Given the lack of differentiation between traits in younger children, perhaps higher-order factors may be more appropriate. For instance, Deyoung (2006) have proposed two higher order factors: *Stability* (made up of *Neuroticism*, *Agreeableness* and *Conscientiousness*) and *Plasticity* (made up of *Openness* and *Extraversion*). Therefore, re-analysis of the data to see if these two higher-order factors emerge is a recommended future avenue for this research. If these higher-order factors do emerge, this would highlight that personality in children may need to be treated differently to personality in adults, and additionally show that the framework under which a test is designed has crucial implications for understanding and interpreting the results.

In Chapter 3, we have replicated the finding that synaesthetes have higher *Openness* traits in a random child sample. In combination with adult findings (Banissy et al., 2013; Chun & Hupé, 2016; Rouw & Scholte, 2016), the link to *Openness* seems quite robust. However, we additionally found differences to the previous literature which may warrant further investigation: are such conflicts the result of differences between children and adults, or

methodological differences? These methodological issues concerned recruitment and verification of synaesthetes and were discussed in detail in Chapter 3. In order to investigate whether these differences are due to age or methodology a random sample of adult synaesthetes would need to be tested on the adult equivalent of the tests we used (BFI-44). One of the most interesting findings from Chapter 3 was that there were diverging personality profiles between the two variants of synaesthesia tested. OLP synaesthetes showing increased *Conscientiousness* traits, and grapheme-colour synaesthetes showing decreased *Extraversion* traits. This finding highlighted that experiences (such as synaesthesia) can influence and shape personality to the point where people who share a particular sensory trait also tend to share a particular personality trait. We can further explore this area to establish the mechanisms that tie cognitive experience and personality development. The two types of synaesthesia studied in the Chapter 3 are similar in nature; they both have grapheme inducers. Therefore, it may be that different concurrents (i.e. the type of resulting synaesthetic associations) can explain differences in the personality profile experienced. Investigation of synaesthesia with the same concurrents but different inducers (e.g., grapheme-colour vs. music-colour) may help disassociate the influences. For example, if music-colour synaesthetes (for whom music evokes colour' see Ward, Huckstep, & Tsakanikos, 2006) show the same personality traits as grapheme-colour synaesthetes (i.e. higher *Openness* and lower *Extraversion*) this would provide evidence of the importance of concurrents.

Another possible element of this puzzle is *synaesthetic dose*. Synaesthetic dose refers to the number of types of synaesthesia a person has with higher “dosages” associated with more complex synaesthetic experiences. Spiller, Jonas, Simner and Jansari (2015) were among the first to note that the number of types of synaesthesia experienced were related to the strength of a particular skill, in their case visual mental imagery. Synaesthetic dose has additionally been shown to correspond to a more autistic-like profile (Ward, Brown, Sherwood, & Simner, 2017). It may therefore be the case that differences in personality profile are associated with synaesthetic dose, rather than different synaesthetic variants per say (see Ward, 2019) for a discussion of ‘dose’ in synaesthesia). A topic of future study might therefore be to investigate whether the personality profile is associated with synaesthetes having more or fewer synaesthesias (see Ward, 2019). Indeed brief inspection of five child synaesthetes in my testing cohort who had *both* forms of synaesthesia suggest they may have had the most extreme personality scores. Although

they were excluded from our analysis of Chapter 3, given our focus on comparing/separating sub-types (and also given their small cohort size), these five synaesthetes showed some of the highest scores (e.g., a post-hoc analysis using Definitional-BFI-44-C scores, and including this group resulted in an odds ratio of 8.04 or a change in the odds of 704% associated with being a synaesthete with both types of synaesthesia and having a higher *Openness* score). In summary, we have shown that synaesthetes tend to have a particular personality profile related to the type of synaesthesia they have. But we have suggested, following Ward (2019), that one avenue for future research might focus on investigating to what extent different personality profiles differ with dose versus type of synaesthesia.

Numerical cognition: What we have learnt?

The second half of the thesis examined numerical cognition. Chapter 4 investigated how environmentally learned colour associations may affect numerical cognition, whilst Chapter 5 specifically focuses on grapheme-colour synaesthetes. In Chapter 4, we investigated whether colour could help aid numerical cognition in non-synaesthetes. We presented our large cohort of children with two tests of numerical cognition: numerosity (‘number sense’) and mathematics. We also measured to what degree the children had internalised the colours two pedagogical tools: *Numicon* (a ‘maths manipulative’ that associates colours to magnitudes) and *Numberjacks* (another maths tool, which associates numbers to Arabic numerals). We investigated whether children who used the colours from these tools would also would score better in our numerical cognition tests. We proposed a dual-coding model in which colour is associated with numbers at two different levels. At the first level, colour may be associated with the magnitude of a number, which would be reflected by children having number colours that match the colours of the *Numicon* classroom tool. At the second level, colour may be associated with the symbolic representation of a number, in which case we expected children to have colours for numbers that match the colours of the *Numberjacks* TV show. We hypothesised that children who encoded colours for magnitude would show improvements in numerosity (perhaps feeding also into improved mathematics), while children who encoded colours at the symbolic/numeral level would show improvements in mathematics only. We found that only children who had internalised magnitude-based colours (i.e. from *Numicon*) did

indeed show improvements in numerosity, as predicted. We found no other significant effects, and therefore updated our model to reflect these findings: dual-coding magnitude leads to improvements in numerosity but not mathematics, and dual-coding symbols does not lead to benefits in either. Overall, Chapter 4 shows that coloured stimuli in the environment does play a role in numerical cognition, but only for magnitude and not for symbols.

Finally, Chapter 5 returned to synaesthesia and investigated whether grapheme-colour synaesthetes showed the same number cognition profile as magnitude-encoding non-synaesthetes. We found the same pattern of results as before: benefits associated with numerosity, but no difference between synaesthetes and non-synaesthetes in mathematics. Importantly, we looked at synaesthetes who only experienced colour for letters and compared them to synaesthetes who had associations for numbers. We found that *letter-only synaesthetes* showed no benefits in numerosity but *number synaesthetes* did. We suggested in Chapter 5 that benefits associated with mathematical skills may be a direct result of synaesthetic experiences. In this case, having extra colour information associated with numbers may lead to improved numerical cognition via dual-coding. Here we were able to apply the same dual-coding model we proposed in Chapter 4. Overall, these results suggest that synaesthesia itself may play a role in the cognitive profile associated with synaesthesia. Synaesthetes with particular associations may experience benefits as a direct result of having their associations. Taken together, Chapters 4 and 5 propose and provide evidence for a model explaining how colour associations may aid in numerical cognition, and at what level of numerical processing the colours may be encoded.

Future research in Synaesthesia and Numerical Cognition

In Chapters 4 and 5 we looked at numerical cognition and the role that colour plays. One of the key findings from Chapters 4 and 5 was that children who have internalised coloured numbers, either from grapheme-colour synaesthesia or from an educational tool, did not show any improvements in their mathematical ability, despite improvements in numerosity. This is unexpected because numerosity independently is associated with increased mathematics performance both in the literature (Halberda et al., 2008) and in our sample specifically (See Chapter 4). Furthermore, synaesthetes have a number of *other* abilities, too, which also correlate with improved mathematics, such as improved

memory abilities (Rothen et al., 2012) and greater IQ (Rouw & Scholte, 2016). All these skills (enhanced numerosity, memory, IQ) have all been correlated with increased mathematical performance (Alloway & Passolunghi, 2011; Halberda et al., 2008; Lynn & Mikk, 2009). One question to address, therefore, is whether synaesthetes perform *worse* in mathematics than would be predicted by their scores on mathematics correlates. In other words, our null effect (no difference between synaesthetes and peers) could be re-conceptualised as a deficit, given the high levels we would otherwise anticipate from their other abilities. This question could be addressed by testing synaesthetes on tasks associated with increased mathematical performance and using these scores to predict how well synaesthetes *should* score in mathematics. If actual mathematical performance is significantly worse than predicted, this would imply that coloured numbers can interfere with mathematics, and that synaesthetes may need additional help to reach their full potential in mathematics.

Another important avenue in numerical cognition and synaesthesia research is to extend the current findings to different synaesthesia types. In Chapter 5, we chose to focus on grapheme-colour synaesthesia only in order to test the model we presented in Chapter 4 in non-synaesthetes. Since this hypothesis pertained to dual-coded colour, we focused on grapheme-colour synaesthesia specifically. However, it will be important to take these findings and investigate different types of synaesthesia to better determine the mechanisms at play. For example, if synaesthetes with non-grapheme inducers *also* experience benefits in numerosity, we would have to revisit and revise our “dual-coding” account. Similarly, future research should consider individuals with multiple types of synaesthesia (see above for our discussion on ‘dose’).

One potential methodological issue may have affected the results of the current studies and warrants further investigation. In our experiments in Chapters 4 and 5, we have used a short non-symbolic dot comparison task as our test of numerosity. There has been some research into what other abilities may be needed in order to complete the task aside from the approximate number system (ANS; i.e. the underlying mechanism behind numerosity; Merkley, Thompson, & Scerif, 2016). One ability needed to complete the dot comparison task is inhibition control. This is the ability to suppress salient but task-irrelevant information (Merkley et al., 2016). A common example of this is in a classic Stroop experiment (Stroop, 1935) in which the participant has to name the font-colour of a word, and avoid reading the word itself (for example reading the word “red” written in yellow

ink). Successfully completing this task requires good inhibition control to ignore the salient but task-irrelevant word. There is a similar degree of ‘strooping’ in the numerosity dot task: here, trials with more numerous dots can be ‘congruent’ if they also take up a bigger space on the screen (i.e., dots are larger), or ‘incongruent trials’ if they take up a smaller space on the screen (i.e., dots are smaller). Participants must ignore the salient but task-irrelevant size of the dots and focus on their number. In our task, it may have been the case that in order to do well on incongruent trials you need a good level of inhibition control.

We discuss this possibility in Chapter 5, but in order to further investigate numerosity whilst accounting for these potential confounds, further research should use a greater array of numerosity tasks. These tasks should include non-symbolic and symbolic numerosity tasks, and account for inhibition control (e.g., using the NEPSY-II inhibition task; Brooks, Sherman, & Strauss, 2009). We predict that synaesthetes should remain significantly better in numerosity ability after accounting for inhibition control. It might be expected that synaesthetes would have significantly better inhibition control itself because synaesthetes often have to ignore salient yet task-irrelevant synaesthetic information in everyday life (i.e., ignoring colour information whilst reading, writing, and dealing with numbers on a day-to-day basis). However, Rouw, van Driel, Knip, and Ridderinkhof (2013) found no differences between synaesthetes and non-synaesthetes in their inhibition control abilities using a Stroop paradigm. Finally, although we would predict that synaesthetes should show benefits in numerosity dependent on synaesthesia type (i.e., advantages for *number synaesthetes*, but not *letter-only synaesthetes*), there is no *a priori* reason to believe that these types of synaesthetes should differ in inhibition control.

Here I have summarised the key findings from each of the studies in this thesis. Above I have also highlighted the next steps in this research: further investigation as to whether synaesthetes experience difficulties in mathematical ability, and investigating a potential alternative account for our numerosity findings concerning inhibition control. In the next section, I explore another avenue of future research that pertains to synaesthesia in children more generally.

Other future directions

Thus far I have summarised the key findings from this thesis and discussed the next directions within each of the two key domains examined in the current thesis. Here I take a step back and look at a key future direction in synaesthesia research in children more generally. This key remaining question is: *how does synaesthesia develop alongside the synaesthete?* For instance, how do synaesthetic associations change as synaesthesia becomes more fixed with age? In OLP synaesthesia, for example, do child synaesthetes experience their personifications as growing up along with them, or are they fixed as their consistent adult-like associations already? In addressing these questions we may be able to provide a richer picture of what synaesthetes are like. We may additionally gain insights into the development of other related processes in children. For example, in children with OLP synaesthesia we may gain insights into how children see other individuals or characters in order to learn more about social development. Furthermore, by learning more about the development of the synaesthesia we may be able to investigate the small number of children with synaesthesia who have problems related to their synaesthesia. Occasionally, parents get in touch with synaesthesia researchers because they have a child having trouble with synaesthesia. For example, a child with OLP may experience a phobia of a letter with a “bad” personality leading to problems in school. As these children are not typical of the synaesthetic experience these experiences are typically missed when investigating group-wise comparisons of synaesthetes and non-synaesthetes in their wellbeing or cognition.

To investigate the development of synaesthesia, we must first obtain the correct type of data. The tests we used to identify OLP and grapheme-colour synaesthetes were primarily *discrimination tests* - their main objective was to distinguish between synaesthetes and non-synaesthetes. Given this main objective, these tasks simplified the detail-rich synaesthetic experience in order to easily distinguish between the two groups. For example, our OLP test for children simply measured three levels of valence (positive, neutral, negative) rather than detailed personalities. In order to answer these emerging questions about how synaesthesia develops, it will be essential to create *experiential* tasks, which gather wider information about an individual’s internal experience, and which can be administered once synaesthetes are identified. One avenue for OLP synaesthesia may be to use a character-generating game similar to The Sims (Electronic Arts Inc., 2019). In this game, children build characters including their appearance, age, personality traits, job aspirations, and relationships to other characters. This game could be adapted such

that synaesthetes could describe grapheme representations other than human (e.g., personification of the letter A) and different stages of life (child, teenager, adult); they would be able to specify relationships (e.g., mother/daughter, friends, enemies) with other graphemes, and they would be able to specify key personality traits. For each grapheme, a number of data points would therefore be subsequently available about each of these areas. This data could be used to measure in either a longitudinal or a cross-sectional study the development of OLP over the lifespan. Building characters in this way at multiple time points throughout their life may provide an age-appropriate and detail-rich way of qualitatively (and perhaps potentially quantitatively) assessing the development of personifications over the course of childhood. Alternatively, in testing a cross-sectional group of children, adolescents, and adult synaesthetes, we may be able to see patterns emerging which help us determine how synaesthetic associations develop over time.

Here I have argued the need for experiential tasks to gather detail-rich synaesthetic associations. These tasks are not the same as discrimination tasks, but these are often confounded in synaesthesia research. In the next section, I will discuss the insights I have gained throughout the process of conducting this research and what recommendations I might make for future researchers who plan to test large cohorts of children.

Designing and running tests for a large cohort of children

In this thesis, I have reported on the results of a large-scale, multi-year research project. In order to complete the study within the time and budget constraints, this research by necessity required testing large cohorts of children at once rather than individually in series. Testing in large groups, rather than testing individually, may have an impact on the results obtained. Here I argue that this group approach had drawbacks but also important benefits, both of which may have had direct impacts on the conclusions that we can draw from our results.

There are a few drawbacks that are important to consider when testing large cohorts of children in groups. Firstly, it is much more difficult to control for individual differences in ability needed to perform each type of task. In a typical UK classroom there are approximately 25-30 students, which vary in ability level (Department for Education,

2014). Therefore, when giving a task in a classroom environment, it is difficult to ascertain which children are struggling with any given task. Even if teachers are able to assist researchers to point out which children may struggle, it is still difficult to provide the level of aid every child needs. Additionally, children who otherwise would accept help in a one-on-one testing environment may be less likely to self-identify as needing help when in a classroom environment due to the added social pressure of being in class (Ryan, Patrick, & Shim, 2005). In order to facilitate testing in a group environment, it might be possible to use digital tools. For example, the paper version of our OLP diagnostic led to difficulties with legibility, children missing certain letters out, or picking more than one face per letter. We therefore digitized the task. Not only did this remove these difficulties (the program would not advance until the task was completed properly), but also this served to expedite data coding. Digital tools also offer potential aids; for example, it would be possible to add a help button to tell the researcher which children need help without the child asking out loud.

A second key drawback is that work produced in a group environment is likely to be more influenced by the individuals around each child, due to copying, collaborating, or distraction. This may be more detrimental to some types of tests over others. For example, we distinguished above between *experiential tests* (which gather information about an individual's experience, like our personality testing) and *discrimination tests* (which identify individuals with a specific trait, like our synaesthesia diagnostic), but there are also *knowledge based tests* (which establish how much an individual knows on a topic; like our maths testing). For knowledge-based tasks in particular, testing in a class environment may actually be beneficial, because these are widely used in schools in any case (e.g., a spelling test) and typically administered in classrooms (Earle, 2019). This means that testing in this way may be a more naturalistic way of gathering information. In turn, it may give a more accurate assessment of children's ability levels. In contrast, the classroom setting is likely to be more detrimental for experiential and discrimination tests, as these aim to gather the internal experience of the child and are less associated with classroom settings.

For researchers interested in testing large numbers of children, it is critical to plan carefully for a few reasons. First, if there is a choice, a different setting may be preferable (group vs individual testing) depending on the type of task of interest, the type of data, and crucially the number of individuals to test. Second, in order to conduct discrimination

or experiential tasks as part of a group rather than an individual testing environment there may be difficulties in finding existing well-validated tasks. When planning our studies, we found that tests often assumed a one-on-one testing environment (e.g., numerical tests provided in testing batteries like numerical operations in the WIAT, or in personality assessments such as the Berkeley Puppet Interview), or were only available for older children aged 12 and above (e.g., Big Five Questionnaire-Children; Muris et al., 2005). Therefore, throughout the thesis I have either adapted existing measures (e.g., the BFI-44-A) or developed new tests (e.g., my test of mathematics). Lastly, I advise it is important to conduct pilot testing, as when testing in groups there is no opportunity to learn about any problems until a substantial number of data points have been collected. We were unable to do extensive piloting in the current study due to external constraints (e.g., finding enough schools to meet our sample-size requirements) and many early practical difficulties had to be resolved after testing had started (e.g., switching from pencil-and-paper testing to tablet-testing for the Pictorial-BFI-10-C personality test reported in Chapter 2). Fortunately, these differences did not affect the comparisons of interest in this thesis (e.g., see Chapter 2, where paper and electronic testing of personality produced equivalent outcomes).

Conclusions

In the sections above, I have summarised the findings from this thesis, suggested future directions based on these findings, and elaborated on the design characteristics for research with large cohorts of children. I now return to the central question in this thesis: *What are child synaesthetes like beyond their synaesthesia?* To address this question, I tested large groups of child OLP and grapheme-colour synaesthetes. I have found that child synaesthetes experience higher levels of the *Openness* trait, related to intelligence and creativity. I found that grapheme-colour synaesthetes experience lower levels of *Extraversion* traits, and OLP synaesthete children experience higher levels of *Conscientiousness*. (These conclusions were possible because I also showed, in Chapter 2, that children as young as 8 years can self-report personality within the Big Five personality model, and that their parents can give similarly robust judgements.) In numerical cognition, I found that non-synaesthetes who internalised colours of educational number tools, and grapheme-colour synaesthetes alike, showed a similar

pattern of results. They significantly outperform controls in a dot numerosity task, but show no similar advantages in mathematics. In this discussion, I have also highlighted some of the key avenues for future research in both of these fields, and have also highlighted some of the key considerations involved in testing large cohorts of children. Overall, with this thesis, we are now indeed one step closer to understanding what synaesthetes are like, given my study of the largest, randomly sampled cohort of child synaesthetes to date.

References

- Ablow, J. C., & Measelle, J. R. (1993). *Berkeley Puppet Interview: Administration and scoring system manuals*. Berkeley: University of California.
- Alloway, T. P., & Passolunghi, M. C. (2011). The relationship between working memory, IQ, and mathematical skills in children. *Learning and Individual Differences*, 21, 133–137. <https://doi.org/10.1016/j.lindif.2010.09.013>
- Amin, M., Olu-Lafe, O., Claessen, L. E., Sobczak-Edmans, M., Ward, J., Williams, A. L., & Sagiv, N. (2011). Understanding grapheme personification: A social synaesthesia? *Journal of Neuropsychology*, 5(2), 255–282. <https://doi.org/10.1111/j.1748-6653.2011.02016.x>
- Anobile, G., Stievano, P., & Burr, D. C. (2013). Visual sustained attention and numerosity sensitivity correlate with math achievement in children. *Journal of Experimental Child Psychology*, 116(2), 380–391. <https://doi.org/10.1016/j.jecp.2013.06.006>
- Asher, J., Lamb, J. A., Brocklebank, D., Cazier, J.-B. B., Maestrini, E., Addis, L., ... Monaco, A. P. (2009). A whole-genome scan and fine-mapping linkage study of auditory-visual synesthesia reveals evidence of linkage to chromosomes 2q24, 5q33, 6p12, and 12p12. *American Journal of Human Genetics*, 84(2), 279–285. <https://doi.org/10.1016/j.ajhg.2009.01.012>
- Auyeung, B., Wheelwright, S. J., Allison, C., Atkinson, M., Samarawickrema, N., & Baron-Cohen, S. (2009). The children's empathy quotient and systemizing quotient: Sex differences in typical development and in autism spectrum conditions. *Journal of Autism and Developmental Disorders*, 39(11), 1509–1521. <https://doi.org/10.1007/s10803-009-0772-x>
- Banissy, M. J., Holle, H., Cassell, J., Annett, L., Tsakanikos, E., Walsh, V., ... Ward, J. (2013). Personality traits in people with synaesthesia: Do synaesthetes have an atypical personality profile? *Personality and Individual Differences*, 54(7), 828–831. <https://doi.org/10.1016/j.paid.2012.12.018>
- Barbaranelli, C., Caprara, G. V., Rabasca, A., & Pastorelli, C. (2003). A questionnaire for measuring the Big Five in late childhood. *Personality and Individual Differences*, 34, 645–664. [https://doi.org/10.1016/S0191-8869\(02\)00051-X](https://doi.org/10.1016/S0191-8869(02)00051-X)

- Barnett, K. J., & Newell, F. N. (2008). Synaesthesia is associated with enhanced, self-rated visual imagery. *Conscious Cogn*, 17(3), 1032–1039. <https://doi.org/10.1016/j.concog.2007.05.011>
- Baron-Cohen, S., Harrison, J., Goldstein, L. H., & Wyke, M. A. (1993). Coloured speech perception: Is synaesthesia what happens when modularity breaks down? *Perception*, 22(4), 419–426. <https://doi.org/10.1068/p220419>
- Baron-Cohen, S., Wyke, M. A., & Binnie, C. (1987). Hearing words and seeing colours: An experimental investigation of a case of synaesthesia. *Perception*, 16(6), 761–767. <https://doi.org/10.1068/p160761>
- Barth, H., Kanwisher, N., & Spelke, E. S. (2003). The construction of large number representations in adults The construction of large number representations in adults. *Cognition*, 86(3), 201–221. [https://doi.org/10.1016/S0010-0277\(02\)00178-6](https://doi.org/10.1016/S0010-0277(02)00178-6)
- Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1), 1–48. <https://doi.org/10.18637/jss.v067.i01>
- Benet-Martínez, V., & John, O. P. (1998). Los Cinco Grandes Across Cultures and Ethnic Groups: Multitrait Multimethod Analyses of the Big Five in Spanish and English. *Journal of Personality and Social Psychology*, 75(3), 729–750. <https://doi.org/10.1037/0022-3514.75.3.729>
- Berlin, B., & Kay, P. (1969). *Basic color terms: Their universality and evolution*. Berkeley: University of California.: University of California Press.
- Berteletti, I., Hubbard, E. M., & Zorzi, M. (2010). Implicit versus explicit interference effects in a number-color synesthete. *Cortex*, 46(2), 170–177. <https://doi.org/10.1016/j.cortex.2008.12.009>
- Birmaher, B., Brent, D., Chiappetta, L., Bridge, J., Monga, S., & Baugher, M. (1999). Psychometric properties of the Screen for Child Anxiety Related Emotional Disorders (SCARED): a replication study. *Journal of the American Academy of Child and Adolescent Psychiatry*. <https://doi.org/10.1097/00004583-199910000-00011>
- Birmaher, B., Khetarpal, S., Brent, D., Cully, M., Balach, L., Kaufman, J., & Neer, S. M.

- (1997). The Screen for Child Anxiety Related Emotional Disorders (SCARED): scale construction and psychometric characteristics. *Journal of the American Academy of Child and Adolescent Psychiatry*. <https://doi.org/10.1097/00004583-199704000-00018>
- Bor, D., Rothen, N., Schwartzman, D. J., Clayton, S., & Seth, A. K. (2014). Adults Can Be Trained to Acquire Synesthetic Experiences. *Scientific Reports*, 4(7089). <https://doi.org/10.1038/srep07089>
- Brooks, B. L., Sherman, E. M. S., & Strauss, E. (2009). Child Neuropsychology NEPSY-II: A Developmental Neuropsychological Assessment, Second Edition. *A Developmental Neuropsychological Assessment*, 16(1), 80–101. <https://doi.org/10.1080/09297040903146966>
- Carbonneau, K. J., Marley, S. C., & Selig, J. P. (2013). A meta-analysis of the efficacy of teaching mathematics with concrete manipulatives. *Journal of Educational Psychology*, 105(2), 380–400. <https://doi.org/10.1037/a0031084>
- Carmichael, D. A., Down, M. P., Shillcock, R. C., Eagleman, D. M., & Simner, J. (2015). Validating a standardised test battery for synesthesia: Does the Synesthesia Battery reliably detect synesthesia? *Consciousness and Cognition*, 33, 375–385. <https://doi.org/10.1016/j.concog.2015.02.001>
- Carmichael, D. A., Smees, R., Shillcock, R. C., & Simner, J. (2018). Is there is a burden attached to synaesthesia? Health screening of synaesthetes in the general population. *British Journal of Psychology*. <https://doi.org/10.1111/bjop.12354>
- Caspi, A., Roberts, B. W., & Shiner, R. (2005). PERSONALITY DEVELOPMENT: Stability and Change. *Annu. Rev. Psychol*, 56, 453–484. <https://doi.org/10.1146/annurev.psych.55.090902.141913>
- Caspi, A., & Shiner, R. (2006). Personality development. In N. Eisenberg, R. Damon, & R. M. Lerner (Eds.), *Handbook of child psychology: Social, emotional, and personality development* (pp. 300–365). Hoboken, NJ, US: John Wiley & Sons, Inc.
- Chen, Q., & Li, J. (2014). Association between individual differences in non-symbolic number acuity and math performance: A meta-analysis. *Acta Psychologica*, 148, 163–172. <https://doi.org/10.1016/j.actpsy.2014.01.016>

- Chun, C. A., & Hupé, J. M. (2016). Are synesthetes exceptional beyond their synesthetic associations? A systematic comparison of creativity, personality, cognition, and mental imagery in synesthetes and controls. *British Journal of Psychology (London, England : 1953)*, 107(3), 397–418. <https://doi.org/10.1111/bjop.12146>
- Churches, R. (2016). *Closing the gap: test and learn*. Retrieved from https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/495580/closing_the_gap_test_and_learn_full_report.pdf
- Clark, J. M., & Paivio, A. (1991). Dual coding theory and education. *Educational Psychology Review*, 3(3), 149–210. <https://doi.org/10.1007/BF01320076>
- Clause-May, T., Vappula, H., & Ruddock, G. (2004). *Progress in Mathematics 6*. GL Assessment.
- Clayton, S., & Gilmore, C. (2015). Inhibition in dot comparison tasks. *ZDM Mathematics Education*, 47, 759–770. <https://doi.org/10.1007/s11858-014-0655-2>
- Clements, D. H., & McMillen, S. (1996). Rethinking “Concrete” manipulatives. *Teaching Children Mathematics*, 2(5), 270–279.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. (2nd ed.). New York: Academic Press.
- Costa, P. T., & McCrae, R. R. (1987). Neuroticism, somatic complaints, and disease: is the bark worse than the bite? *Journal of Personality*, 55(2), 299–316. <https://doi.org/https://doi.org/10.1111/j.1467-6494.1987.tb00438.x>
- Costa, P. T., & McCrae, R. R. (1992). Normal Personality Assessment in Clinical Practice: The NEO Personality Inventory. *Psychological Assessment*, 4(1), 5–13. <https://doi.org/10.1037/1040-3590.4.1.5>
- Credé, M., Harms, P. D., Niehorster, S., & Gaye-Valentine, A. (2012). An Evaluation of the Consequences of Using Short Measures of the Big Five Personality Traits. *Journal of Personality and Social Psychology*, 102(4), 874–888. <https://doi.org/10.1037/a0027403>
- Danner, D., Aichholzer, J., & Rammstedt, B. (2015). Acquiescence in personality questionnaires: Relevance, domain specificity, and stability. *Journal of Research in Personality*, 57, 119–130. <https://doi.org/10.1016/j.jrp.2015.05.004>

- Day, J., & Lockwood, J. (2008). Multisensory mathematics for wave 3 intervention. In *Doncaster Council*.
- Dehaene, S. (2001). Précis of “ The number sense .” *Mind and Language*, 16(1), 16–36.
<https://doi.org/10.1111/j.1468-0017.1986.tb00086.x>
- Dehaene, S., & Cohen, L. (1995). Towards an anatomical and functional model of number processing. *Mathematical Cognition*, 1(1), 83–120.
- Dehaene, S., & Cohen, L. (1997). Cerebral pathways for calculation: Double dissociation between rote verbal and quantitative knowledge of arithmetic. *Cortex*, 33(2), 219–250.
 Retrieved from <https://www.sciencedirect.com/science/article/pii/S0010945208700029>
- Department for Education. (2014). *Class Size and Education in England*. Retrieved from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/183364/DFE-RR169.pdf
- Deroy, O., & Spence, C. (2013). Are we all born synaesthetic? Examining the neonatal synaesthesia hypothesis. *Neurosci Biobehav Rev*, 37(7), 1240–1253.
<https://doi.org/10.1016/j.neubiorev.2013.04.001>
- Devon Primary Maths Team. (2006). *An image of number: The use of numicon in mainstream classrooms*. Exeter: Devon County Council.
- DeWind, N. K., & Brannon, E. M. (2012). Malleability of the approximate number system: effects of feedback and training. *Frontiers in Human Neuroscience*, 6, 1–10.
<https://doi.org/10.3389/fnhum.2012.00068>
- Deyoung, C. G. (2006). Higher-Order Factors of the Big Five in a Multi-Informant Sample. *Journal of Personality and Social Psychology*, 91(6), 1138–1151.
<https://doi.org/10.1037/0022-3514.91.6.1138>
- Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology*, 5(July), 1–17. <https://doi.org/10.3389/fpsyg.2014.00781>
- Dittmar, A. (2009). *Synaesthesia: A "golden Thread" Through Life?* (A. Dittmar, Ed.). Verlag Die Blaue Eule.
- Domino, G. (1989). Synesthesia and Creativity in Fine Arts Students: An Empirical Look.

- Creativity Research Journal*, 2(1–2), 17–29.
<https://doi.org/10.1080/10400418909534297>
- Eagleman, D. M., Kagan, A. D., Nelson, S. S., Sagaram, D., & Sarma, A. K. (2007). A standardized test battery for the study of synesthesia. *Journal of Neuroscience Methods*, 159(1), 139–145. <https://doi.org/10.1016/j.jneumeth.2006.07.012>
- Earle, S. (2019). *Assessment in the Primary Classroom: Principles and practice*. London: Learning Matters.
- Education Leeds. (2008). *Multi-sensory approach to the teaching and learning of mathematics: Pilot project 2005*.
- Eisinga, R., Grotenhuis, M. Te, & Pelzer, B. (2013). The reliability of a two-item scale: Pearson, Cronbach, or Spearman-Brown? *International Journal of Public Health*, 58(4), 637–642. <https://doi.org/10.1007/s00038-012-0416-3>
- Electronic Arts Inc. (2019). The Sims. Retrieved from <https://www.ea.com/games/the-sims>
- Elias, L. J., Saucier, D. M., Hardie, C., & Sarty, G. E. (2003). Dissociating semantic and perceptual components of synaesthesia: behavioural and functional neuroanatomical investigations. *Brain Research. Cognitive Brain Research*, 16(2), 232–237. [https://doi.org/10.1016/S0926-6410\(02\)00278-1](https://doi.org/10.1016/S0926-6410(02)00278-1)
- Ellis, C. (2006). BBC - CBeebies - Numberjacks. Retrieved May 29, 2017, from <http://www.bbc.co.uk/programmes/b006mhcr>
- Ester, M., Kriegel, H.-P., Jorg, S., & Xu, X. (1996). A Density-Based Clustering Algorithms for Discovering Clusters. *Kdd-96 Proceedings*, 96(34), 226–231. <https://doi.org/10.1016/B978-044452701-1.00067-3>
- Ewan, C., & Mair, C. (2002). Wiltshire Pilot Project - Numicon (March - July 2001). *Down Syndrome News and Update*, 2(1), 12–13.
- Feigenson, L., Dehaene, S., & Spelke, E. S. (2004). Core systems of number. *Trends in Cognitive Sciences*, 8(7), 307–314. <https://doi.org/10.1016/j.tics.2004.05.002>
- Field, A. P., Miles, J., & Field, Z. (2012). *Discovering Statistics using R*. London: SAGE.
- Ganz Cooney, J., Mottisett, L., & Morrisett, L. (2019). Sesame Street. Retrieved from

<http://www.sesamestreet.org/>

- García, L. F., Aluja, A., García, Ó., & Cuevas, L. (2005). Is openness to experience an independent personality dimension? Convergent and discriminant validity of the openness domain and its NEO-PI-R facets. *Journal of Individual Differences*, 26(3), 132-138.
- Gertner, L., Arend, I., & Henik, A. (2013). Numerical synesthesia is more than just a symbol-induced phenomenon. *Frontiers in Psychology*, 4(November), 1–4. <https://doi.org/10.3389/fpsyg.2013.00860>
- Gibson, B. S., Radvansky, G. A., Johnson, A. C., & McNerney, M. W. (2012). Grapheme–color synesthesia can enhance immediate memory without disrupting the encoding of relational cues. *Psychonomic Bulletin & Review*, 19(6), 1172–1177. <https://doi.org/10.3758/s13423-012-0306-y>
- Goldberg, L. R. (1990). An Alternative Description of Personality - the Big-5 Factor Structure. *Journal of Personality and Social Psychology*, 59(6), 1216–1229. <https://doi.org/10.1037//0022-3514.59.6.1216>
- Goldsmith, H. H., Buss, A. H., Plomin, R., Rothbart, M. K., Thomas, A., Chess, S., ... Mccall, R. B. (1987). Roundtable : What Is Temperament ? Four Approaches. *Child Development*, 58(2), 505–529. Retrieved from <http://www.jstor.org/stable/1130527>
- Goodman, R. (1997). The Strengths and Difficulties Questionnaire: a research note. *Journal of Child Psychology and Psychiatry*, 38(5), 581–586. <https://doi.org/10.1111/j.1469-7610.1997.tb01545.x>
- Gorard, S., See, B. H., & Siddiqui, N. (2017). *The trials of evidence-based education: The promises, opportunities and problems of trials in education*. Taylor & Francis.
- Gosling, S. D., Rentfrow, P. J., & Swann, W. B. (2003). A very brief measure of the Big-Five personality domains. *Journal of Research in Personality*, 37, 504–528. [https://doi.org/10.1016/S0092-6566\(03\)00046-1](https://doi.org/10.1016/S0092-6566(03)00046-1)
- Green, J. A. K., & Goswami, U. (2008). Synesthesia and number cognition in children. *Cognition*, 106(1), 463–473. <https://doi.org/10.1016/j.cognition.2007.01.013>
- Grieve, R. (2012). The Role of Personality, Psychopathy, and Previous Experience with Assessment in Intentions to Fake in Psychological Testing. *Current Psychology*,

- 31(4), 414–422. <https://doi.org/10.1007/s12144-012-9158-x>
- Gross, V. C., Nearing, S., Caldwell-Harris, C. L., & Cronin-Golomb, A. (2011). Superior encoding enhances recall in color-graphemic synesthesia. *Perception*, 40(2), 196–208. <https://doi.org/10.1068/p6647>
- Grossenbacher, P. G., & Lovelace, C. T. (2001). Mechanisms of synesthesia: Cognitive and physiological constraints. *Trends in Cognitive Sciences*, 5(1), 36–41. [https://doi.org/10.1016/S1364-6613\(00\)01571-0](https://doi.org/10.1016/S1364-6613(00)01571-0)
- Halberda, J., Mazocco, M. M. M., & Feigenson, L. (2008). Individual differences in non-verbal number acuity correlate with maths achievement. *Nature*, 455(7213), 665–668. <https://doi.org/10.1038/nature07246>
- Halverson, C. F., Havill, V. L., Deal, J., Baker, S. R., Victor, J. B., Pavlopoulos, V., ... Wen, L. (2003). Personality Structure as Derived from Parental Ratings of Free Descriptions of Children: The Inventory of Child Individual Differences. *Journal of Personality*, 71(6), 995–1026. <https://doi.org/10.1111/1467-6494.7106005>
- Hänggi, J., Wotruba, D., & Jäncke, L. (2011). Globally altered structural brain network topology in grapheme-color synesthesia. *Journal of Neuroscience*, 31(15), 5816–5828. <https://doi.org/10.1523/JNEUROSCI.0964-10.2011>
- Havlik, A. M., Carmichael, D. A., & Simner, J. (2015). Do sequence-space synaesthetes have better spatial imagery skills? Yes, but there are individual differences. *Cognitive Processing*, 16(3), 245–253. <https://doi.org/10.1007/s10339-015-0657-1>
- Hopwood, C. J., & Donnellan, M. B. (2010). How should the internal structure of personality inventories be evaluated? *Personality and Social Psychology Review*, 14(3), 332–346. <https://doi.org/10.1177/1088868310361240>
- Hughes, J. E. A., Ipser, A., & Simner, J. (2019). The MULTISENSE test for sequence-personality synaesthesia. *In Preparation*.
- Hurewitz, F., Papafragou, A., Gleitman, L., & Gelman, R. (2006). Asymmetries in the acquisition of numbers and quantifiers. *Language Learning and Development*, 2(2), 77–96. https://doi.org/10.1207/s15473341l1d0202_1
- Hutcheson, G., & Sofroniou, N. (1999). *The multivariate social scientist: Introductory statistics using generalized linear models*. London: SAGE Publications.

- International Color Consortium ®. (2004). Specification ICC.1:2004-10 (Profile version 4.2.0.0). Image technology colour management — Architecture, profile format, and data structure.
- Janik McErlean, A. B., & Banissy, M. J. (2016). Examining the relationship between schizotypy and self-reported visual imagery vividness in grapheme-color synaesthesia. *Frontiers in psychology*, 7, 131.
- John, O. P. (2009). Berkeley Personality Lab. Retrieved from <https://www.ocf.berkeley.edu/~johnlab/bfi.php>
- John, O. P., Caspi, A., Robins, R. W., Moffitt, T. E., & Stouthamer-Loeber, M. (1994). The “Little Five”: Exploring the Nomological Network of the Five-Factor Model of Personality in Adolescent Boys. *Child Development*, 65(1), 160–178. <https://doi.org/10.1111/j.1467-8624.1994.tb00742.x>
- John, O. P., Donahue, E. M., & Kentle, R. L. (1991). *The big five inventory—versions 4a and 54*. University of California, Berkeley, Institute of Personality and Social Research, Berkeley, CA.
- John, O. P., Naumann, L. P., & Soto, C. J. (2008). Paradigm Shift to the Integrative Big Five Trait Taxonomy. In L. Pervin & O. P. John (Eds.), *Handbook of personality: Theory and research* (pp. 114–158).
- John, O. P., & Srivastava, S. (1999). The Big Five trait taxonomy: History, Measurement, and Theoretical Perspectives. In L. Pervin & O. John (Eds.), *Handbook of personality: Theory and research (2nd ed.)*. <https://doi.org/citeulike-article-id:3488537>
- Kadosh, R. C., Sagiv, N., Linden, D. E. J., Robertson, L. C., Elinger, G., & Henik, A. (2005). When blue is larger than red: Colors influence numerical cognition in synesthesia. *Journal of Cognitive Neuroscience*, 17(11), 1766–1773. <https://doi.org/10.1162/089892905774589181>
- Kennis, M., Rademaker, A. R., & Geuze, E. (2013). Neural correlates of personality: An integrative review. *Neuroscience and Biobehavioral Reviews*, 37, 73–95. <https://doi.org/10.1016/j.neubiorev.2012.10.012>
- Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition

- ratings for 30,000 English words. *Behavior Research Methods*, 44, 978–990.
<https://doi.org/10.3758/s13428-012-0210-4>
- Kusnir, F., & Thut, G. (2012). Formation of automatic letter-colour associations in non-synaesthetes through likelihood manipulation of letter-colour pairings. *Neuropsychologia*, 50, 3641–3652.
<https://doi.org/10.1016/j.neuropsychologia.2012.09.032>
- Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, 82(13), 1–26.
<https://doi.org/10.18637/jss.v082.i13>
- Lee, K., & Kang, S. . (2002). Arithmetic operation and working memory: Differential suppression in dual tasks. *Cognition*, 83(3). [https://doi.org/10.1016/S0010-0277\(02\)00010-0](https://doi.org/10.1016/S0010-0277(02)00010-0)
- Lee, M., & Wagenmakers, E. (2014). *Bayesian Cognitive Modeling: A practical Course*. Cambridge: Cambridge University Press.
- Lipton, J. S., & Spelke, E. S. (2004). Discrimination of large and small numerosities by human infants. *Infancy*, 5(3), 271–290. https://doi.org/10.1207/s15327078in0503_2
- Lorenzo-Seva, U., & ten Berge, J. M. F. (2006). Tucker’s congruence coefficient as a meaningful index of factor similarity. *Methodology*, 2(2), 57–64.
<https://doi.org/10.1027/1614-2241.2.2.57>
- Luo, L., & O’Leary, D. D. M. (2005). Axon retraction and degeneration in development and disease. *Annual Review of Neuroscience*, 28, 127–156.
<https://doi.org/10.1146/annurev.neuro.28.061604.135632>
- Lynn, R., & Mikk, J. (2009). National IQs predict educational attainment in math, reading and science across 56 nations. *Intelligence*, 37(3), 305–310.
<https://doi.org/10.1016/j.intell.2009.01.002>
- Mackiewicz, M., & Ciecuch, J. (2016). Pictorial Personality Traits Questionnaire for Children (PPTQ-C)-a new measure of children’s personality traits. *Frontiers in Psychology*, 7, 1–11. <https://doi.org/10.3389/fpsyg.2016.00498>
- Macleod, C. M., & Dunbar, K. N. (1988). Training and Stroop-Like Interference: Evidence for a Continuum of Automaticity. *Journal of Experimental Psychology*

- Learning Memory and Cognition*, 14(1), 126. <https://doi.org/10.1037/0278-7393.14.1.126>
- Malone, S. A., Pritchard, V. E., Heron-Delaney, M., Burgoyne, K., Lervåg, A., & Hulme, C. (2019). The relationship between numerosity discrimination and arithmetic skill reflects the approximate number system and cannot be explained by inhibitory control. *Journal of Experimental Child Psychology*, 184, 220–231. <https://doi.org/10.1016/j.jecp.2019.02.009>
- Mankin, J. L., & Simner, J. (2017). A is for apple: The role of letter-word associations in the development of grapheme-colour synaesthesia. *Multisensory Research*, 30(3–5), 409–446. <https://doi.org/10.1163/22134808-00002554>
- Mares, M.-L., & Pan, Z. (2013). Effects of Sesame Street: A meta-analysis of children's learning in 15 countries. *Journal of Applied Developmental Psychology*, 34(3), 140–151. <https://doi.org/10.1016/J.APPDEV.2013.01.001>
- Markey, P. M., Markey, C. N., & Tinsley, B. J. (2004). Children's Behavioral Manifestations of the Five-Factor Model of Personality. *Personality and Social Psychology Bulletin*, 30(4), 423–432. <https://doi.org/10.1177/0146167203261886>
- Markey, P. M., Markey, C. N., Tinsley, B. J., & Ericksen, A. J. (2002). A preliminary validation of preadolescents' self-reports using the five-factor model of personality. *Journal of Research in Personality*, 36(2), 173–181. <https://doi.org/10.1006/jrpe.2001.2341>
- Maurer, D. (1993). Neonatal synesthesia: Implications for the processing of speech and faces. In *Developmental neurocognition: Speech and face processing in the first year of life* (pp. 109–124). Dordrecht: Springer.
- Maurer, D., Gibson, L. C., & Spector, F. (2013). Synesthesia in infants and very young children. In J. Simner & E. M. Hubbard (Eds.), *The Oxford Handbook of Synesthesia* (pp. 46–53). Oxford: Oxford University Press.
- McCrae, R. R., & Costa, P. T. (1987). Validation of the Five-Factor Model of Personality Across Instruments and Observers. *Journal of Personality and Social Psychology*, 52(1), 81–90. <https://doi.org/10.1037/0022-3514.52.1.81>
- McNeil, N. M., Uttal, D. H., Jarvin, L., & Sternberg, R. J. (2009). Should you show me

- the money? Concrete objects both hurt and help performance on mathematics problems. *Learning and Instruction*, 19(2), 171–184. <https://doi.org/10.1016/j.learninstruc.2008.03.005>
- Measelle, J. R., John, O. P., Ablow, J. C., Cowan, P. A., & Cowan, C. P. (2005). Can Children Provide Coherent, Stable, and Valid Self-Reports on the Big Five Dimensions? A Longitudinal Study From Ages 5 to 7. *Journal of Personality and Social Psychology*, 89(1), 90–106. <https://doi.org/10.1037/0022-3514.89.1.90>
- Mednick, S. A. (1968). The remote associates test. *The Journal of Creative Behavior*, 2(3), 213–214. <https://doi.org/10.1002/j.2162-6057.1968.tb00104.x>
- Meier, B., & Rothen, N. (2009). Training grapheme-colour associations produces a synaesthetic Stroop effect, but not a conditioned synaesthetic response. *Neuropsychologia*, 47(4), 1208–1211. <https://doi.org/10.1016/j.neuropsychologia.2009.01.009>
- Meier, B., & Rothen, N. (2013a). Grapheme-color synaesthesia is associated with a distinct cognitive style. *Front Psychol*, 4, 632. <https://doi.org/10.3389/fpsyg.2013.00632>
- Meier, B., & Rothen, N. (2013b). Synesthesia and Memory. In J. Simner & E. M. Hubbard (Eds.), *The Oxford Handbook of Synesthesia* (pp. 692–707). Oxford: Oxford University Press.
- Meier, B., Rothen, N., & Walter, S. (2014). Developmental aspects of synaesthesia across the adult lifespan. *Frontiers in Human Neuroscience*, 8(March), 129. <https://doi.org/10.3389/fnhum.2014.00129>
- Meisenberg, G., & Williams, A. (2008). Are acquiescent and extreme response styles related to low intelligence and education? *Personality and Individual Differences*, 44(7), 1539–1550. <https://doi.org/10.1016/j.paid.2008.01.010>
- Melchers, M. C., Li, M., Haas, B. W., Reuter, M., Bischoff, L., & Montag, C. (2016). Similar Personality Patterns Are Associated with Empathy in Four Different Countries. *Frontiers in Psychology*, 7(March), 1–12. <https://doi.org/10.3389/fpsyg.2016.00290>
- Merkley, R., Thompson, J., & Scerif, G. (2016). Of huge mice and tiny elephants:

- Exploring the relationship between inhibitory processes and preschool math skills. *Frontiers in Psychology*. <https://doi.org/10.3389/fpsyg.2015.01903>
- Mervielde, I., & De Fruyt, F. (1999). Construction of the Hierarchical Personality Inventory for Children (HiPIC). In I. Mervielde, I. Deary, F. De Fruyt, & F. Ostendorf (Eds.), *Personality psychology in Europe. Proceedings of the Eight European Conference on Personality Psychology* (pp. 107–127). Tilburg: Tilburg University Press.
- Mills, C. B., Metzger, S. R., Foster, C. A., Valentine-Gresk, M. N., & Ricketts, S. (2009). Development of color- grapheme synesthesia and its effect on mathematical operations. *Perception*, 38(4), 591–695. <https://doi.org/10.1068/p6109>
- Morey, R., Rouder, J., & Jamil, T. (2015). Package “BayesFactor.” Retrieved August 5, 2019, from alvarestech.com website: <ftp://alvarestech.com/pub/plan/R/web/packages/BayesFactor/BayesFactor.pdf>
- Munroe, R. (2010). Colour Survey Results. Retrieved from <https://blog.xkcd.com/2010/05/03/color-survey-results/>
- Muris, P., Bos, A. E. R., Mayer, B., Verkade, R., Thewissen, V., & Dell’Avvento, V. (2009). Relations among behavioral inhibition, Big Five personality factors, and anxiety disorder symptoms in non-clinical children. *Personality and Individual Differences*, 46(4), 525–529. <https://doi.org/10.1016/j.paid.2008.12.003>
- Muris, P., Meesters, C., & Dideren, R. (2005). Psychometric properties of the Big Five Questionnaire for Children (BFQ-C) in a Dutch sample of young adolescents. *Personality and Individual Differences*, 38, 1757–1769. <https://doi.org/10.1016/j.paid.2004.11.018>
- Neville, H. J. (1995). Developmental specificity in neurocognitive development in humans. In M. S. Gazzaniga (Ed.), *The cognitive neurosciences* (pp. 219–231). Cambridge: MIT Press.
- Nye, J., Buckley, S., & Bird, G. (2005). Evaluating the Numicon system as a tool for teaching number skills to children with Down syndrome. *Down Syndrome News and Update*, 5(1), 2–13.
- O’Donnell, L. (2009). The Wechsler intelligence scale for children—fourth edition. In J.

- A. Naglieri & S. Goldstein (Eds.), *Practitioner's guide to assessing intelligence and achievement* (pp. 153–190). Hoboken, NJ, US: John Wiley & Sons, Inc.
- Oxford University Press. (2018). Numicon, Primary School Maths Resources. Retrieved February 2, 2018, from <https://global.oup.com/education/content/primary/series/numicon/?region=uk>
- Paivio, A. (1969). Mental imagery in associative learning and memory. *Psychological Review*, 76(3), 241–263. <https://doi.org/10.1037/h0027272>
- Piaget, J. (1965). The stages of the intellectual development of the child. In *Educational psychology in context: Readings for future teachers* (pp. 98–106).
- Pritchard, J., Rothen, N., Coolbear, D., & Ward, J. (2013). Enhanced associative memory for colour (but not shape or location) in synaesthesia. *Cognition*, 127(2), 230–234. <https://doi.org/10.1016/j.cognition.2012.12.012>
- R Core Team. (2016). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.
- Radvansky, G. A., Gibson, B. S., & McNerney, M. W. (2011). Synesthesia and Memory: Color Congruency, von Restorff, and False Memory Effects. *Journal of Experimental Psychology: Learning Memory and Cognition*, 37(1), 219–229. <https://doi.org/10.1037/a0021329>
- Ramachandran, V. S., & Azoulay, S. (2006). Synesthetically Induced Colors Evoke Apparent-Motion Perception. *Perception*, 35(11), 1557–1560. <https://doi.org/10.1068/p5565>
- Rammstedt, B., & Farmer, R. F. (2013). The impact of acquiescence on the evaluation of personality structure. *Psychological Assessment*, 25(4), 1137–1145. <https://doi.org/10.1037/a0033323>
- Rammstedt, B., & John, O. P. (2007). Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. *Journal of Research in Personality*, 41, 20–212. <https://doi.org/10.1016/j.jrp.2006.02.001>
- Riccelli, R., Toschi, N., Nigro, S., Terracciano, A., & Passamonti, L. (2017). Surface-based morphometry reveals the neuroanatomical basis of the five-factor model of personality. *Social Cognitive and Affective Neuroscience*, 671–684.

<https://doi.org/10.1093/scan/nsw175>

- Rich, A. N., Bradshaw, J. L., & Mattingley, J. B. (2005). A systematic, large-scale study of synaesthesia: Implications for the role of early experience in lexical-colour associations. *Cognition*, 98(1), 53–84. <https://doi.org/10.1016/j.cognition.2004.11.003>
- Rinaldi, L. J., Smees, R., Alvarez, J., & Simner, J. (2019). Do the colors of educational number-tools improve children’s mathematics and numerosity? *Child Development*, *in press*.
- Rinaldi, L. J., Smees, R., Carmichael, D. A., & Simner, J. (2019a). Big Five Personality Instruments for Parents and Children 6+ years: The Pictorial BFI-10-C; the Definitional BFI-44-c, and the BFI-44-parent. *In Preparation*.
- Rinaldi, L. J., Smees, R., Carmichael, D. A., & Simner, J. (2019b). What is the personality profile of a child synaesthete? *Frontiers in Biosciences*, *in press*.
- Ripley, B., & Venables, W. (2016). Nnet: feed-forward neural networks and multinomial log-linear models.
- Roberts, B. W., & DelVecchio, W. F. (2000). The rank-order consistency of personality from childhood to old age: A quantitative review of longitudinal studies. *Psychological Bulletin*, 126(1), 3–25. <https://doi.org/10.1037//0033-2909.126.1.3>
- Robertson, L. C., & Sagiv, N. (2004). *Synaesthesia: perspectives from cognitive neuroscience* (L. C. Robertson & N. Sagiv, Eds.). Oxford University Press.
- Rothen, N., & Meier, B. (2010a). Grapheme-colour synaesthesia yields an ordinary rather than extraordinary memory advantage: Evidence from a group study. *Memory*, 18(3), 258–264. <https://doi.org/10.1080/09658210903527308>
- Rothen, N., & Meier, B. (2010b). Higher prevalence of synaesthesia in art students. *Perception*, 39(5), 718–720. <https://doi.org/10.1068/p6680>
- Rothen, N., Meier, B., & Ward, J. (2012). Enhanced memory ability: Insights from synaesthesia. *Neuroscience and Biobehavioral Reviews*, 36(8), 1952–1963. <https://doi.org/10.1016/j.neubiorev.2012.05.004>
- Rothen, N., Seth, A. K., Witzel, C., & Ward, J. (2013). Diagnosing synaesthesia with

- online colour pickers: Maximising sensitivity and specificity. *Journal of Neuroscience Methods*, 215(1), 156–160. <https://doi.org/10.1016/j.jneumeth.2013.02.009>
- Rothen, N., Wantz, A.-L., & Meier, B. (2011). Training synaesthesia. *Perception*, 40, 1248–1250. <https://doi.org/10.1068/p6984>
- Rouder, J. N., & Morey, R. D. (2012). Multivariate Behavioral Research Default Bayes Factors for Model Selection in Regression. *Multivariate Behavioral Research*, 47(February 2013), 877–903. <https://doi.org/10.1080/00273171.2012.734737>
- Rouw, R., & Scholte, H. S. (2007). Increased structural connectivity in grapheme-color synesthesia. *Nature Neuroscience*, 10(6), 792–797. <https://doi.org/10.1038/nn1906>
- Rouw, R., & Scholte, H. S. (2016). Personality and cognitive profiles of a general synesthetic trait. *Neuropsychologia*, 88, 35–48. <https://doi.org/10.1016/j.neuropsychologia.2016.01.006>
- Rouw, R., Scholte, H. S., & Colizoli, O. (2011). Brain areas involved in synaesthesia: a review. *J Neuropsychol*, 5(2), 214–242. <https://doi.org/10.1111/j.1748-6653.2011.02006.x>
- Rouw, R., van Driel, J., Knip, K., & Richard Ridderinkhof, K. (2013). Executive functions in synesthesia. *Conscious Cogn*, 22(1), 184–202. <https://doi.org/10.1016/j.concog.2012.11.008>
- Ryan, A. M., Patrick, H., & Shim, S.-O. (2005). Differential Profiles of Students Identified by Their Teacher as Having Avoidant, Appropriate, or Dependent Help-Seeking Tendencies in the Classroom. *Journal of Educational Psychology*, 97(2), 275–285. <https://doi.org/10.1037/0022-0663.97.2.275>
- Simner, J. (2012). Defining synaesthesia: A response to two excellent commentaries. *British Journal of Psychology*, 103(1), 24–27. <https://doi.org/10.1111/j.2044-8295.2011.02059.x>
- Simner, J. (2019). *Synaesthesia. A Very Short Introduction*. Oxford: Oxford University Press.
- Simner, J., Alvarez, J., Rinaldi, L. J., Smees, R., & Carmichael, D. A. (2019). The MULTISENSE sequence-personality diagnostic for children. *In Preparation*.

- Simner, J., Alvarez, J., Smees, R., Rinaldi, L. J., & Carmichael, D. A. (2019). The MULTISENSE test for grapheme-colour synaesthesia in children. *In Preparation*.
- Simner, J., & Bain, A. E. (2013). A longitudinal study of grapheme-color synesthesia in childhood: 6/7 years to 10/11 years. *Frontiers in Human Neuroscience*, 7(November), 603. <https://doi.org/10.3389/fnhum.2013.00603>
- Simner, J., & Bain, A. E. (2018). Do children with grapheme-colour synaesthesia show cognitive benefits? *British Journal of Psychology*, 109(1), 118–136. <https://doi.org/10.1111/bjop.12248>
- Simner, J., Gartner, O., & Taylor, M. D. (2011). Cross-modal personality attributions in synaesthetes and non-synaesthetes. *Journal of Neuropsychology*, 5(2), 283–301. <https://doi.org/10.1111/j.1748-6653.2011.02009.x>
- Simner, J., Glover, L., & Mowat, A. (2006). Linguistic Determinants of Word Colouring in Grapheme-Colour Synaesthesia. *Cortex*, 42(2), 281–289. [https://doi.org/10.1016/S0010-9452\(08\)70353-8](https://doi.org/10.1016/S0010-9452(08)70353-8)
- Simner, J., Harrold, J., Creed, H., Monro, L., & Foulkes, L. (2009). Early detection of markers for synaesthesia in childhood populations. *Brain*, 132(1), 57–64. <https://doi.org/10.1093/brain/awn292>
- Simner, J., & Holenstein, E. (2007). Ordinal linguistic personification as a variant of synesthesia. *Journal of Cognitive Neuroscience*, 19(4), 694–703. <https://doi.org/10.1162/jocn.2007.19.4.694>
- Simner, J., & Hubbard, E. M. (2013). *Oxford Handbook of Synesthesia* (Julia Simner & E. M. Hubbard, Eds.). Oxford: OUP Oxford.
- Simner, J., & Logie, R. H. (2007). Synaesthetic consistency spans decades in a lexical-gustatory synaesthete. *Neurocase*, 13(5), 358–365. <https://doi.org/10.1080/13554790701851502>
- Simner, J., Mulvenna, C., Sagiv, N., Tsakanikos, E., Witherby, S. A. S. A., Fraser, C., ... Ward, J. (2006). Synaesthesia: The prevalence of atypical cross-modal experiences. *Perception*, 35(8), 1024–1033. <https://doi.org/10.1068/p5469>
- Simner, J., Rehme, M. K., Carmichael, D. A., Bastin, M. E., Sprooten, E., McIntosh, A. M., ... Zedler, M. (2016). Social responsiveness to inanimate entities: Altered white

- matter in a ‘social synaesthesia.’ *Neuropsychologia*, 91(7), 282–289. <https://doi.org/10.1016/j.neuropsychologia.2016.08.020>
- Simner, J., Rinaldi, L. J., Alvarez, J., Smees, R., Ipser, A., & Carmichael, D. A. (2019). The MULTISENSE grapheme-colour diagnostic for children. *In Preparation*.
- Smees, R., Hughes, J. E. A., Simner, J., & Carmichael, D. A. (2019). Learning in colour: Children with grapheme-colour synaesthesia show cognitive benefits in vocabulary and self-evaluated reading. *Philosophical Transactions of the Royal Society B*, in press.
- Soto, C. J., & John, O. P. (2009). Ten facet scales for the Big Five Inventory: Convergence with NEO PI-R facets, self-peer agreement, and discriminant validity. *Journal of Research in Personality*, 43(1), 84–90. <https://doi.org/10.1016/j.jrp.2008.10.002>
- Soto, C. J., John, O. P., Gosling, S. D., & Potter, J. (2008). The Developmental Psychometrics of Big Five Self-Reports: Acquiescence, Factor Structure, Coherence, and Differentiation From Ages 10 to 20. *Journal of Personality and Social Psychology*, 94(4), 718–737. <https://doi.org/10.1037/0022-3514.94.4.718>
- Soto, C. J., John, O. P., Gosling, S. D., & Potter, J. (2011). Age Differences in Personality Traits From 10 to 65: Big Five Domains and Facets in a Large Cross-Sectional Sample. *Journal of Personality and Social Psychology*, 100(2), 330–348. <https://doi.org/10.1037/a0021717>
- Spiller, M. J., Jonas, C. N., Simner, J., & Jansari, A. (2015). Beyond visual imagery: How modality-specific is enhanced mental imagery in synesthesia? *Consciousness and Cognition*, 31, 73–85. <https://doi.org/10.1016/j.concog.2014.10.010>
- Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, 18(6), 643. <https://doi.org/10.1037/h0054651>
- Tacon, R., Atkinson, R., & Wing, T. (2004). *Learning about numbers with patterns. Using structured visual imagery (Numicon) to teach arithmetic*.
- Taylor, C. (2018). The Reliability of Free School Meal Eligibility as a Measure of Socio-Economic Disadvantage: Evidence from the Millennium Cohort Study in Wales. *British Journal of Educational Studies*, 66(1), 29–51.

<https://doi.org/10.1080/00071005.2017.1330464>

- The national curriculum in England: Key stages 1 and 2 framework document. (2013). Retrieved from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/425601/PRIMARY_national_curriculum.pdf
- Tupes, E. C., & Christal, R. E. (1961). Recurrent Personality Factors Based on Trait Ratings. *Journal of Personality*, 60(2), 225–251. <https://doi.org/10.1111/j.1467-6494.1992.tb00973.x>
- Ward, J. (2019). A Distinct Entity that is an Emergent Feature of Adaptive Neurocognitive Differences. *Philosophical Transactions of the Royal Society B*, in press.
- Ward, J., Brown, P., Sherwood, J., & Simner, J. (2018). An autistic-like profile of attention and perception in synaesthesia. *Cortex*, 107, 121–130. <https://doi.org/10.1016/j.cortex.2017.10.008>
- Ward, J., Hovard, P., Jones, A., & Rothen, N. (2013). Enhanced recognition memory in grapheme-color synaesthesia for different categories of visual stimuli. *Front Psychol*, 4, 762. <https://doi.org/10.3389/fpsyg.2013.00762>
- Ward, J., Huckstep, B., & Tsakanikos, E. (2006). Sound-colour synaesthesia: To what extent does it use cross-modal mechanisms common to us all? *Cortex*, 42(2), 264–280. [https://doi.org/10.1016/S0010-9452\(08\)70352-6](https://doi.org/10.1016/S0010-9452(08)70352-6)
- Ward, J., Ipser, A., Phanvanova, E., Brown, P., Bunte, I., & Simner, J. (2018). The prevalence and cognitive profile of sequence-space synaesthesia. *Consciousness and Cognition*, 61(March), 79–93. <https://doi.org/10.1016/j.concog.2018.03.012>
- Ward, J., Sagiv, N., & Butterworth, B. (2009). The impact of visuo-spatial number forms on simple arithmetic. *Cortex*, 45(10), 1261–1265. <https://doi.org/10.1016/j.cortex.2009.03.017>
- Ward, J., Simner, J., & Auyeung, V. (2005). A comparison of lexical-gustatory and grapheme-colour synaesthesia. *Cognitive Neuropsychology*, 22(1), 28–41. <https://doi.org/10.1080/02643290442000022>
- Ward, J., Thompson-Lake, D., Ely, R., & Kaminski, F. (2008). Synaesthesia, creativity

- and art: What is the link? *British Journal of Psychology*, 99(Pt 1), 127–141.
<https://doi.org/10.1348/000712607X204164>
- Wing, T., & Tacon, R. (2007). Teaching number skills and concepts with Numicon materials. *Down Syndrome Research and Practice*, 12(1), 22–26.
<https://doi.org/10.3104/practice.2018>
- Witthoft, N., & Winawer, J. (2006). Synesthetic colors determined by having colored refrigerator magnets in childhood. *Cortex*, 42(2), 175–183.
[https://doi.org/10.1016/S0010-9452\(08\)70342-3](https://doi.org/10.1016/S0010-9452(08)70342-3)
- Witthoft, N., Winawer, J., & Eagleman, D. M. (2015). Prevalence of learned grapheme-color pairings in a large online sample of synesthetes. *PLoS One*, 10(3), e0118996.
<https://doi.org/10.1371/journal.pone.0118996>
- Wong, T. T.-Y., Ho, C. S.-H., & Tang, J. (2016). The relation between ANS and symbolic arithmetic skills: The mediating role of number-numerosity mappings. *Contemporary Educational Psychology*, 46, 208–217.
<https://doi.org/10.1016/J.CEDPSYCH.2016.06.003>
- Worden, P. E. P. ., & Boettcher, W. (1990). Young Children's Acquisition of Alphabet Knowledge. *Journal of Reading Behavior*, 22(3), 277–295.
<https://doi.org/10.1080/10862969009547711>
- Wright, J. C., Huston, A. C., Murphy, K. C., Peters, M. S., Piñon, M., Scantlin, R., ... Kotler, J. (2001). The Relations of Early Television Viewing to School Readiness and Vocabulary of Children from Low-Income Families : The Early Window Project. *Child Development*, 72(5), 1347–1366. <https://doi.org/10.1111/1467-8624.t01-1-00352>
- Xu, F., & Spelke, E. S. (2000). Large number discrimination in 6-month-old infants. *Cognition*, 74(1), 1–11. [https://doi.org/10.1016/S0010-0277\(99\)00066-9](https://doi.org/10.1016/S0010-0277(99)00066-9)
- Yaro, C., & Ward, J. (2007). Searching for Shereshevskii: What is superior about the memory of synaesthetes? *Quarterly Journal of Experimental Psychology*, 60(5), 681–695. <https://doi.org/10.1080/17470210600785208>

Appendices

Appendix A

Full Questionnaire for the Pictorial-BFI-10-C and full verbal instructions

Children were instructed to “look at the children below and read the words saying what the children are like. One side is a different kind of child to the other side. For each one choose which is the most like you. Once you've chosen which side is like you, put a cross in one of the boxes to say whether it's just a bit like you, sometimes like you, mostly like you or completely like you.” The children were told to ask for help if they had any trouble reading the words, and for children completing the questionnaire on paper they were additionally instructed to only choose one box from one side. All children were given an example question which was read out to them. They were instructed “this child hates eating fruit at lunch but this child loves eating fruit at lunch. So what you need to do is decide which child is most like you. When you've chosen you just press one box to say whether it's just a bit, mostly, sometimes or completely like you”.



About You

Look at the children below and read the words saying what the children are like. On one side is a different kind of child to the other side. Choose which side is most like you!

Then, cross whether it's COMPLETELY like you, MOSTLY like you, SOMETIMES like you or JUST A BIT like you.

Only cross one box!

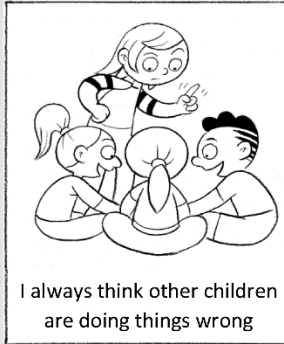
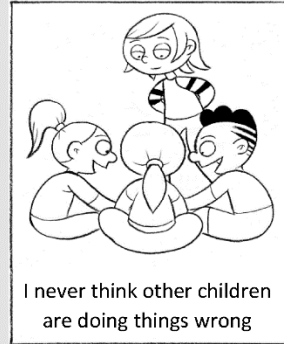
Let's try this one as an example:

		Which one is like you?			
Completely <input type="checkbox"/>	 <p>I hate eating fruit at lunch</p>	 <p>I love eating fruit at lunch</p>	Completely <input type="checkbox"/>		
Mostly <input type="checkbox"/>			Mostly <input type="checkbox"/>		
Sometimes <input type="checkbox"/>			Sometimes <input type="checkbox"/>		
Just a bit <input type="checkbox"/>			Just a bit <input type="checkbox"/>		

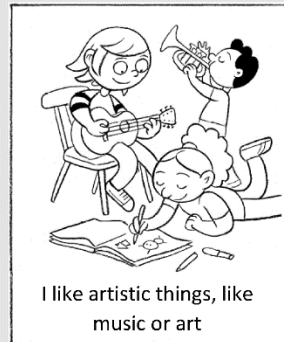
Well done!

Now turn the page and begin.

Which one is like you?

Completely ☐Mostly ☐Sometimes ☐Just a bit ☐☐ Completely☐ Mostly☐ Sometimes☐ Just a bit

Which one is like you?

Completely ☐Mostly ☐Sometimes ☐Just a bit ☐☐ Completely☐ Mostly☐ Sometimes☐ Just a bit

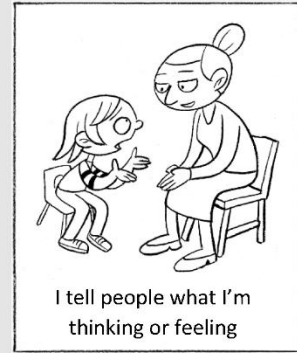
Which one is like you?

Completely ☐Mostly ☐Sometimes ☐Just a bit ☐☐ Completely☐ Mostly☐ Sometimes☐ Just a bit

Which one is like you?

Completely ☐Mostly ☐Sometimes ☐Just a bit ☐☐ Completely☐ Mostly☐ Sometimes☐ Just a bit

Which one is like you?

Completely ☐Mostly ☐Sometimes ☐Just a bit ☐☐ Completely☐ Mostly☐ Sometimes☐ Just a bit

Which one is like you?

Completely ☐Mostly ☐Sometimes ☐Just a bit ☐☐ Completely☐ Mostly☐ Sometimes☐ Just a bit

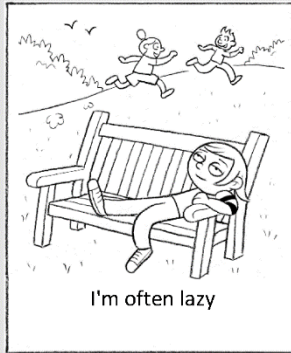
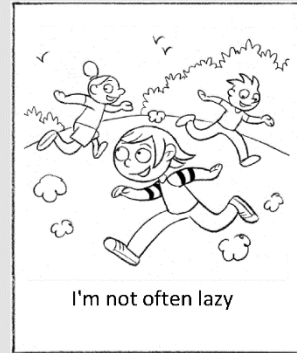
Which one is like you?

Completely ☐Mostly ☐Sometimes ☐Just a bit ☐☐ Completely☐ Mostly☐ Sometimes☐ Just a bit

Which one is like you?

Completely ☐Mostly ☐Sometimes ☐Just a bit ☐☐ Completely☐ Mostly☐ Sometimes☐ Just a bit

Which one is like you?

Completely ☐Mostly ☐Sometimes ☐Just a bit ☐☐ Completely☐ Mostly☐ Sometimes☐ Just a bitCompletely ☐Mostly ☐Sometimes ☐Just a bit ☐

Which one is like you?

☐ Completely☐ Mostly☐ Sometimes☐ Just a bit

Appendix B

Methodology to compute acquiescence-controlled factor scores for the Definitional-BFI-44-C

In order to control for acquiescence it is strongly recommended to compute ipsatized factors. In order to do this for each participant create their mean response and sd for the paired items in the Definitional-BFI-44-C (these items are: BFI1, BFI6, BFI16, BFI21, BFI31, BFI36, BFI2, BFI7, BFI12, BFI17, BFI27, BFI32, BFI37, BFI42, BFI3, BFI8, BFI13, BFI18, BFI23, BFI28, BFI33, BFI43, BFI9, BFI19, BFI24, BFI29, BFI34, BFI39, BFI5, BFI30, BFI35, BFI41).

For each item you compute the new ipsatized item by subtracting the individuals' average from the original item and dividing by the individuals' standard deviation. Note that for reversed items (BFI2, BFI6, BFI8, BFI9, BFI12, BFI18, BFI21, BFI23, BFI24, BFI27, BFI31, BFI34, BFI35, BFI37, BFI41, and BFI43) you will need to first reverse items and then compute ipsatized scores using the reversed items using the same approach.

To compute factors simply use the ipsatized items for a particular factor to compute the mean score for that factor. Listed below are the item-factor correspondence (note that items ending in "r" are reversed).

Openness = BFI5, BFI10, BFI15, BFI20, BFI25, BFI30, BFI35r, BFI40, BFI41r, BFI44

Conscientiousness = BFI3, BFI8r, BFI13, BFI18r, BFI23r, BFI28, BFI33, BFI38, BFI43r

Extraversion = BFI1, BFI6r, BFI11, BFI16, BFI21r, BFI26, BFI31r, BFI36

Agreeableness = BFI2r, BFI7, BFI12r, BFI17, BFI22, BFI27r, BFI32, BFI37r, BFI42

Neuroticism = BFI4, BFI9r, BFI14, BFI19, BFI24r, BFI29, BFI34r, BFI39

If you are using SPSS you can use the provided Syntax:

******Definitional-BFI-44-c syntax file to compute ipsatized items and factor scores*

**This assumes raw data is labelled in SPSS as BFI1, BFI2 etc.*

*** TO REVERSE ITEMS

RECODE

BF12 BF16 BF18 BF19 BF112 BF118 BF121 BF123 BF124 BF127 BF131 BF134 BF135
BF137 BF141 BF143
(1=5) (2=4) (3=3) (4=2) (5=1) INTO BF12r BF16r BF18r BF19r BF112r BF118r
BF121r BF123r BF124r
BF127r BF131r BF134r BF135r BF137r BF141r BF143r.
EXECUTE .

*****IPSATIZING DATA*****

* Compute within-person response means (BFlave) and standard deviations (BF1std).

COMPUTE BFlave = mean(BF11, BF16, BF116, BF121, BF131, BF136, BF12, BF17,
BF112, BF117, BF127, BF132, BF137, BF142, BF13, BF18, BF113, BF118, BF123, BF128,
BF133, BF143, BF19,
BF119, BF124, BF129, BF134, BF139, BF15, BF130, BF135, BF141).
COMPUTE BF1std = sd(BF11, BF16, BF116, BF121, BF131, BF136, BF12, BF17, BF112,
BF117, BF127, BF132, BF137, BF142, BF13, BF18, BF113, BF118, BF123, BF128, BF133,
BF143, BF19, BF119, BF124, BF129, BF134, BF139, BF15, BF130, BF135, BF141).
EXECUTE.

* Compute ipsatizedBFI items (zBFI).

COMPUTE zBF11 = (BF11-BFlave)/BF1std.
COMPUTE zBF12 = (BF12-BFlave)/BF1std.
COMPUTE zBF13 = (BF13-BFlave)/BF1std.
COMPUTE zBF14 = (BF14-BFlave)/BF1std.
COMPUTE zBF15 = (BF15-BFlave)/BF1std.
COMPUTE zBF16 = (BF16-BFlave)/BF1std.
COMPUTE zBF17 = (BF17-BFlave)/BF1std.
COMPUTE zBF18 = (BF18-BFlave)/BF1std.
COMPUTE zBF19 = (BF19-BFlave)/BF1std.
COMPUTE zBF110 = (BF110-BFlave)/BF1std.
COMPUTE zBF111 = (BF111-BFlave)/BF1std.
COMPUTE zBF112 = (BF112-BFlave)/BF1std.
COMPUTE zBF113 = (BF113-BFlave)/BF1std.
COMPUTE zBF114 = (BF114-BFlave)/BF1std.
COMPUTE zBF115 = (BF115-BFlave)/BF1std.
COMPUTE zBF116 = (BF116-BFlave)/BF1std.
COMPUTE zBF117 = (BF117-BFlave)/BF1std.
COMPUTE zBF118 = (BF118-BFlave)/BF1std.
COMPUTE zBF119 = (BF119-BFlave)/BF1std.
COMPUTE zBF120 = (BF120-BFlave)/BF1std.
COMPUTE zBF121 = (BF121-BFlave)/BF1std.
COMPUTE zBF122 = (BF122-BFlave)/BF1std.
COMPUTE zBF123 = (BF123-BFlave)/BF1std.
COMPUTE zBF124 = (BF124-BFlave)/BF1std.
COMPUTE zBF125 = (BF125-BFlave)/BF1std.
COMPUTE zBF126 = (BF126-BFlave)/BF1std.

COMPUTE $zBFI27 = (BFI27 - BFI_{ave}) / BFI_{std}$.
 COMPUTE $zBFI28 = (BFI28 - BFI_{ave}) / BFI_{std}$.
 COMPUTE $zBFI29 = (BFI29 - BFI_{ave}) / BFI_{std}$.
 COMPUTE $zBFI30 = (BFI30 - BFI_{ave}) / BFI_{std}$.
 COMPUTE $zBFI31 = (BFI31 - BFI_{ave}) / BFI_{std}$.
 COMPUTE $zBFI32 = (BFI32 - BFI_{ave}) / BFI_{std}$.
 COMPUTE $zBFI33 = (BFI33 - BFI_{ave}) / BFI_{std}$.
 COMPUTE $zBFI34 = (BFI34 - BFI_{ave}) / BFI_{std}$.
 COMPUTE $zBFI35 = (BFI35 - BFI_{ave}) / BFI_{std}$.
 COMPUTE $zBFI36 = (BFI36 - BFI_{ave}) / BFI_{std}$.
 COMPUTE $zBFI37 = (BFI37 - BFI_{ave}) / BFI_{std}$.
 COMPUTE $zBFI38 = (BFI38 - BFI_{ave}) / BFI_{std}$.
 COMPUTE $zBFI39 = (BFI39 - BFI_{ave}) / BFI_{std}$.
 COMPUTE $zBFI40 = (BFI40 - BFI_{ave}) / BFI_{std}$.
 COMPUTE $zBFI41 = (BFI41 - BFI_{ave}) / BFI_{std}$.
 COMPUTE $zBFI42 = (BFI42 - BFI_{ave}) / BFI_{std}$.
 COMPUTE $zBFI43 = (BFI43 - BFI_{ave}) / BFI_{std}$.
 COMPUTE $zBFI44 = (BFI44 - BFI_{ave}) / BFI_{std}$.

***These items are reverse scored*

COMPUTE $zBFI2r = (BFI2r - BFI_{ave}) / BFI_{std}$.
 COMPUTE $zBFI6r = (BFI6r - BFI_{ave}) / BFI_{std}$.
 COMPUTE $zBFI8r = (BFI8r - BFI_{ave}) / BFI_{std}$.
 COMPUTE $zBFI9r = (BFI9r - BFI_{ave}) / BFI_{std}$.
 COMPUTE $zBFI12r = (BFI12r - BFI_{ave}) / BFI_{std}$.
 COMPUTE $zBFI18r = (BFI18r - BFI_{ave}) / BFI_{std}$.
 COMPUTE $zBFI21r = (BFI21r - BFI_{ave}) / BFI_{std}$.
 COMPUTE $zBFI23r = (BFI23r - BFI_{ave}) / BFI_{std}$.
 COMPUTE $zBFI24r = (BFI24r - BFI_{ave}) / BFI_{std}$.
 COMPUTE $zBFI27r = (BFI27r - BFI_{ave}) / BFI_{std}$.
 COMPUTE $zBFI31r = (BFI31r - BFI_{ave}) / BFI_{std}$.
 COMPUTE $zBFI34r = (BFI34r - BFI_{ave}) / BFI_{std}$.
 COMPUTE $zBFI35r = (BFI34r - BFI_{ave}) / BFI_{std}$.
 COMPUTE $zBFI37r = (BFI37r - BFI_{ave}) / BFI_{std}$.
 COMPUTE $zBFI41r = (BFI41r - BFI_{ave}) / BFI_{std}$.
 COMPUTE $zBFI43r = (BFI43r - BFI_{ave}) / BFI_{std}$.
 EXECUTE.

******FACTORS******
 *

COMPUTE *BFI.e* =
 mean($zBFI1, zBFI6r, zBFI11, zBFI16, zBFI21r, zBFI26, zBFI31r, zBFI36$) .
 VARIABLE LABELS *BFI.e* 'extraversion scale score'.
 EXECUTE .

COMPUTE *BFI.a* =
 mean($zBFI2r, zBFI7, zBFI12r, zBFI17, zBFI22, zBFI27r, zBFI32, zBFI37r, zBFI42$) .
 VARIABLE LABELS *BFI.a* 'agreeableness scale score' .

EXECUTE .

```

COMPUTE                                BFI.c                                =
mean(zBFI3,zBFI8r,zBFI13,zBFI18r,zBFI23r,zBFI28,zBFI33,zBFI38,zBFI43r) .
VARIABLE LABELS BFIc 'conscientiousness scale score' .
EXECUTE .

```

```

COMPUTE                                BFI.n                                =
mean(zBFI4,zBFI9r,zBFI14,zBFI19,zBFI24r,zBFI29,zBFI34r,zBFI39) .
VARIABLE LABELS BFI n 'neuroticism scale score' .
EXECUTE .

```

```

COMPUTE                                BFI.o                                =
mean(zBFI5,zBFI10,zBFI15,zBFI20,zBFI25,zBFI30,zBFI35r,zBFI40,zBFI41r,zBFI44) .
VARIABLE LABELS BFIo 'openness scale score' .
EXECUTE .

```

Appendix C

Methodology to compute acquiescence-controlled factor scores for the Pictorial-BFI-10-C

In order to compute adjusted factor scores for the Pictorial-BFI-10-C you compute for each individual their mean response (Note that this questionnaire does not have matched items unlike the Definitional BFI-44-C so the mean is computed based on all items and the standard deviation is not computed).

Next you compute the adjusted items by taking the original item and subtracting the average from it. Note that you will need to first reverse items 5, 7, 8, 9 and 11.

Finally you compute the factor scores by adding together the adjusted items for each factor shown below:

Openness = BFI3, BFI8r

Conscientiousness = BFI5r, BFI10,

Extraversion = BFI6, BFI11r

Agreeableness = BFI2, BFI7r

Neuroticism = BFI4, BFI9r,

If you are using SPSS you can use the provided Syntax:

******Pictorial-BFI-10 syntax file to compute ipsatized items and factor scores*

**This assumes raw data is labelled in SPSS as pict1, pict2 etc.*

***Please note: this code starts from item 2: pict2 as it assumes pict1 is the test item*

****REVERSE SCORING ITEMS**

COMPUTE pict.5r=9- pict.5.
EXECUTE.

COMPUTE .pict7r= 9- pict.7.
EXECUTE.

COMPUTE pict.8r=9- pict.8.

EXECUTE.

COMPUTE pict.9re=9- pict.9.
EXECUTE.

COMPUTE pict.11r=9- pict.11.
EXECUTE.

******IPSATIZING DATA******

***compute within-person response means (pictave)*

COMPUTE pictave=mean(pict.2, pict.3, pict.4,
pict.5, pict.6, pict.7, pict.8, pict.9, pict.10, pict.11).
EXECUTE.

***compute centered items (zpict)*

COMPUTE zpict.2 = pict.2-pictave.
COMPUTE zpict.3 = pict.3 - pictave.
COMPUTE zpict.4 = pict.4 - pictave.
COMPUTE zpict.5 = pict.5 - pictave.
COMPUTE zpict.6 = pict.6 - pictave.
COMPUTE zpict.7 = pict.7 - pictave.
COMPUTE zpict.8 = pict.8 - pictave.
COMPUTE zpict.9 = pict.9 - pictave.
COMPUTE zpict.10 = pict.10 - pictave.
COMPUTE zpict.11 = pict.11 - pictave.

****reversed items*

COMPUTE zpict.5r = pict.5r- pictave.
COMPUTE zpict.7r = pict.7r- pictave.
COMPUTE zpict.8r = pict.8r- pictave.
COMPUTE zpict.9r = pict.9r- pictave.
COMPUTE zpict.11r = pict.11r- pictave.
EXECUTE.

******FACTORS******

DATASET ACTIVATE DataSet1.
COMPUTE pict.o = zpict.3 + zpict.8r.
EXECUTE.

COMPUTE pict.c = zpict.10 + zpict.5r.
EXECUTE.

COMPUTE $pict.e = zpict.6 + zpict.11r$.
EXECUTE.

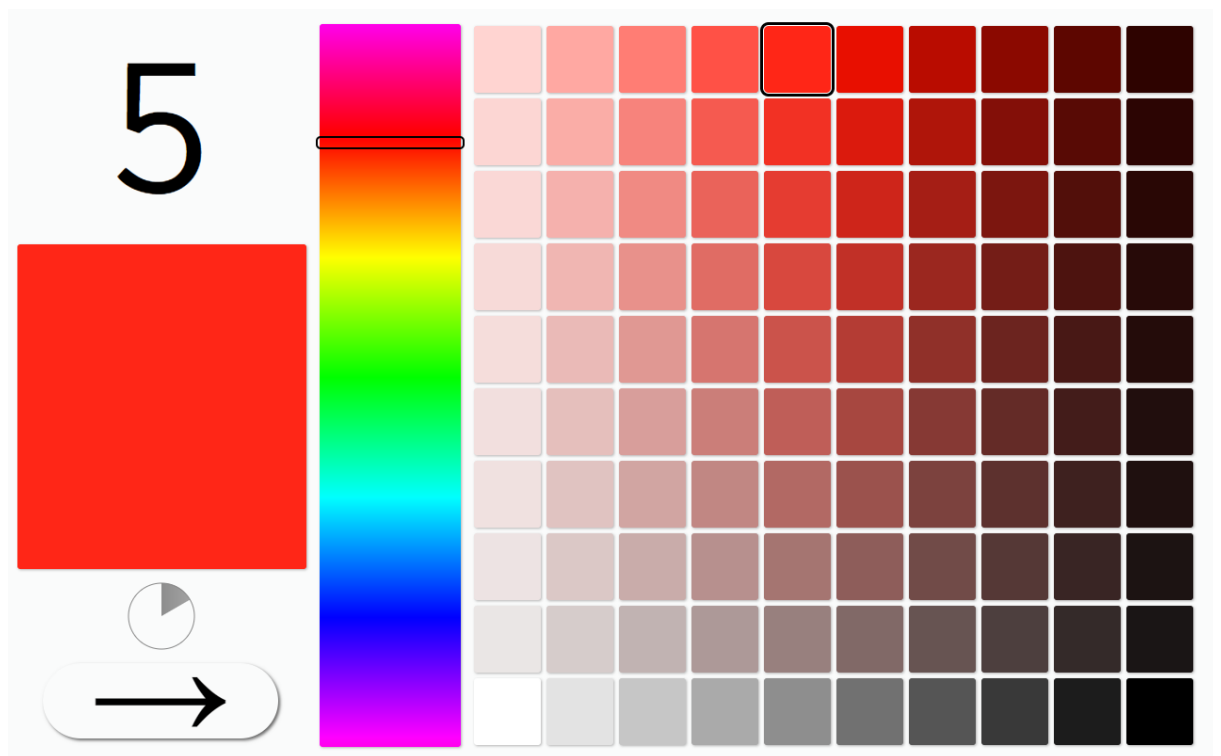
COMPUTE $pict.n = zpict.4 + zpict.9r$.
EXECUTE.

COMPUTE $pict.a = zpict.2 + zpict.7r$.
EXECUTE.

Appendix D

A screenshot of the Grapheme-Colour Diagnostic.

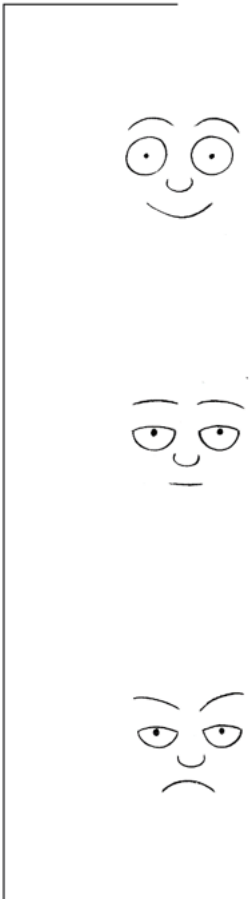
This task was also used in Chapter 4 to identify children using Numicon or Numberjacks number-colour schemas.



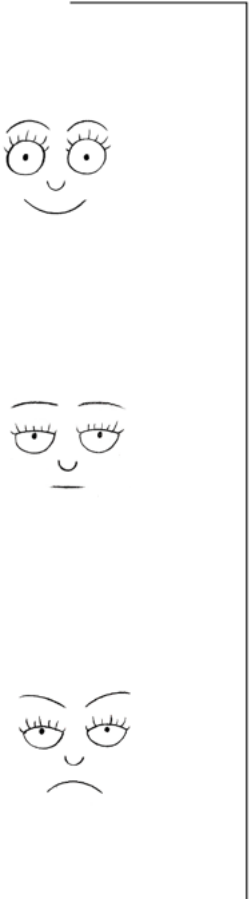
Appendix E

A screenshot of the Ordinal Linguistic Personification Test

Boy faces



Girl faces



G
Q
S
W
F
Z
P
N
J
M
E
T
C
K
O
L
D
H
R
Y
U
A
B
V
I
X

Appendix F

Histograms: To better clarify our groupings

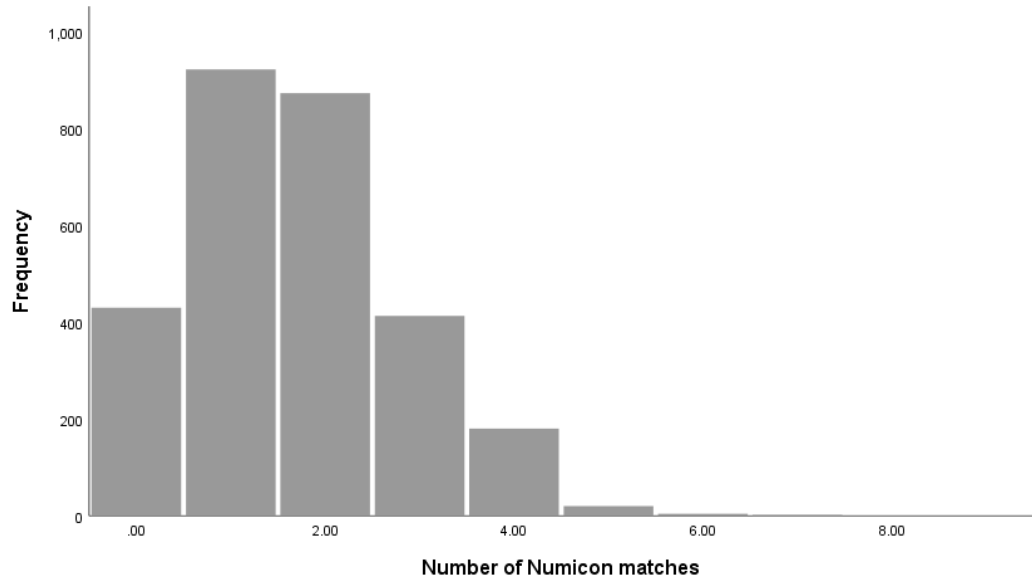


Figure 1. A histogram illustrating the number of Numicon matches for numerosity participants included in the analysed sample.

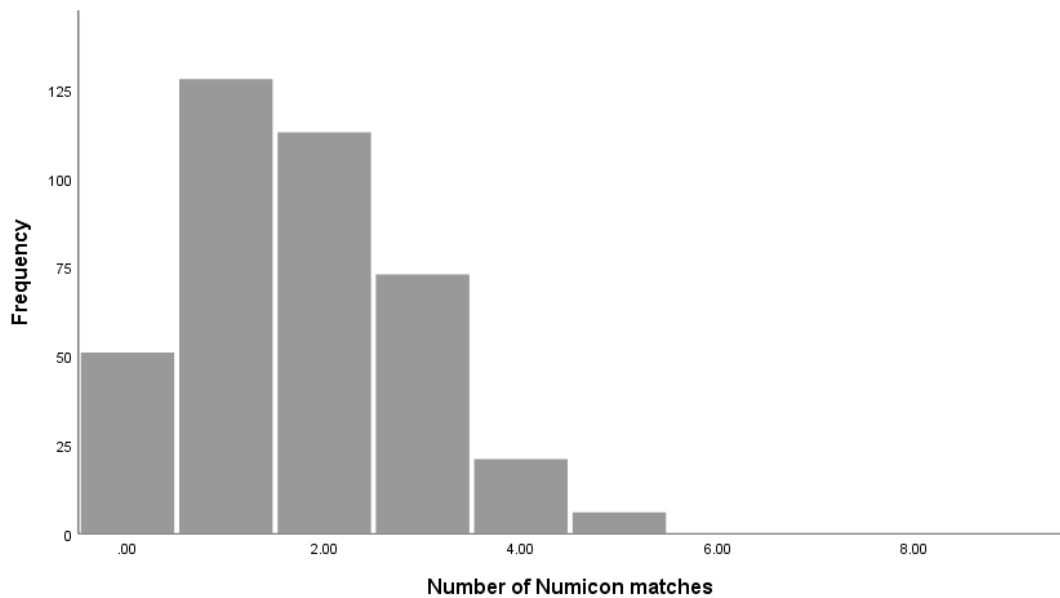


Figure 2. A histogram illustrating the number of Numicon matches for numerosity participants excluded from the analysed sample following current class teacher's indication that they do not use Numicon.

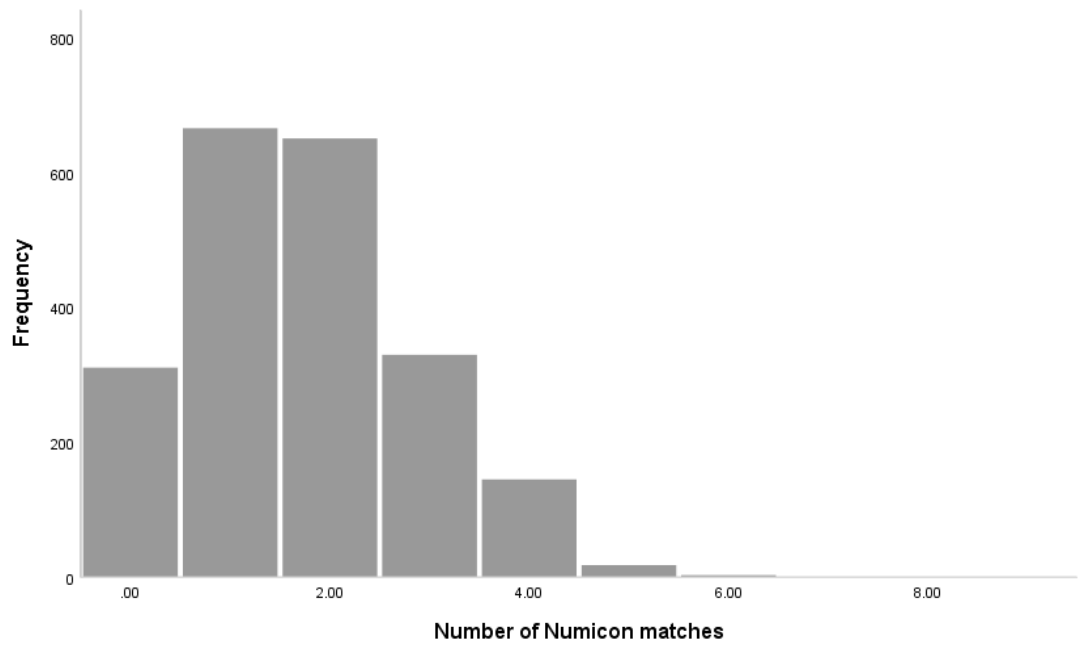


Figure 3. A histogram illustrating the number of Numicon matches for maths participants included in the analysed sample.

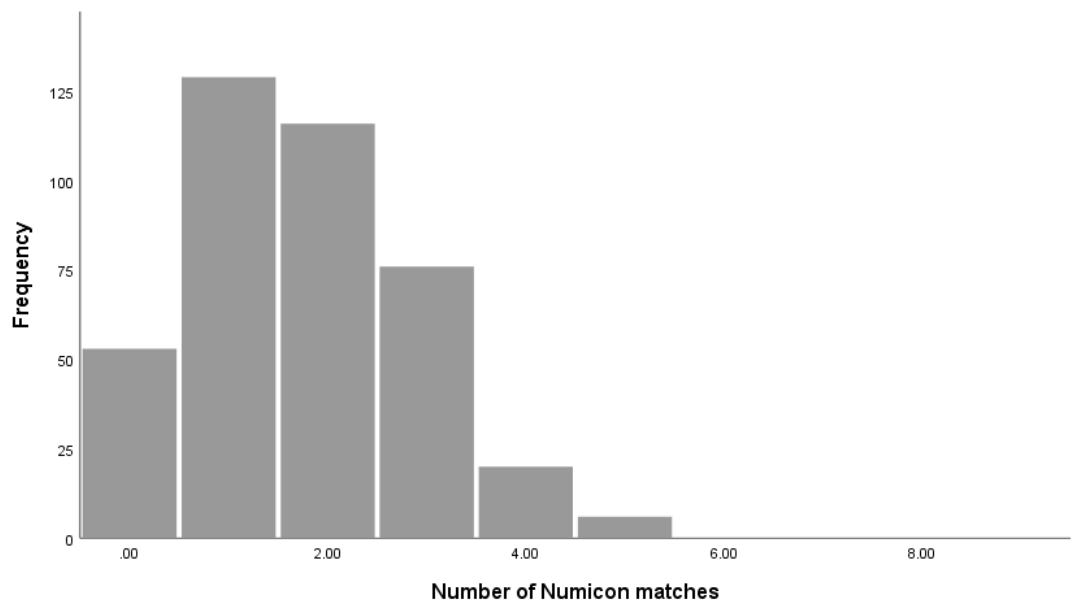


Figure 4. A histogram illustrating the number of Numicon matches for maths participants excluded from the analysed sample following current class teacher's indication that they do not use Numicon.

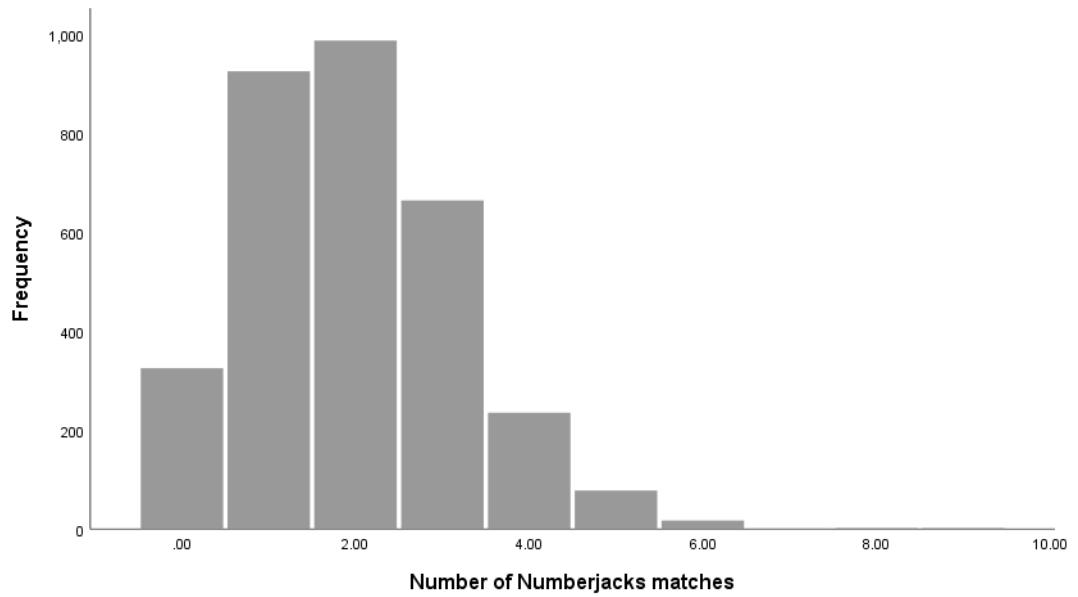


Figure 5. A histogram illustrating the number of Numberjacks matches for numerosity participants included in the analysed sample.

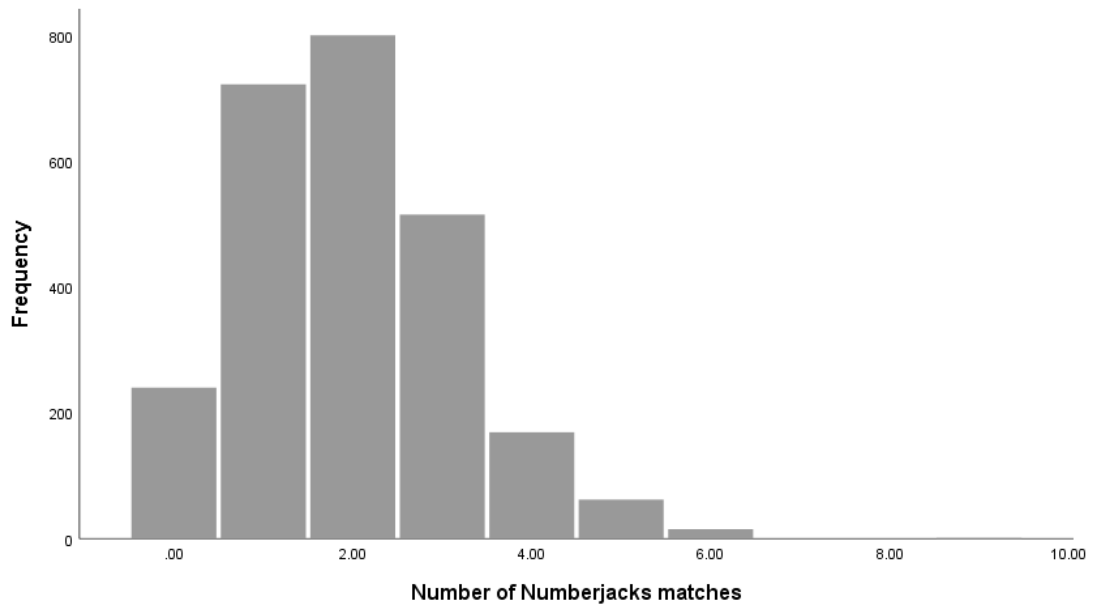


Figure 6. A histogram illustrating the number of Numberjacks matches for maths participants included in the analysed sample.